



# STIC Search Report

## EIC 2100

STIC Database Tracking Number: 169068

**TO: Gwen Liang**  
**Location: RND-3B11**

**Art Unit: 2162**  
**Monday, December 19, 2005**  
**Case Serial Number: 10/752,432**

**From: Lance Sealey**  
**Location: EIC 2100**  
**RND-4B11**  
**Phone: 571-272-8666**  
**Lance.Sealey@uspto.gov**

### Search Notes

Dear Gwen,

Here are the results of your search request.  
Please let me know if you have any questions.

Lance

Green, Shirelle

172155

From: Liang, Gwen  
Sent: Monday, November 21, 2005 2:06 PM  
To: STIC-EIC2100  
Subject: Database Search Request

(92)

Requester:  
GWEN LIANG (TC2100)

Art Unit:  
2162

Employee Number:  
79180

Office Location:  
RND 3B-11

Phone\_Number:  
x24038

Mailbox Number:

Case serial number:  
10/752,432 (PCT/US04/00168)

Class / Subclass(es):

Earliest Priority Filing Date:  
1-8-2003

Format preferred for results:  
Paper

Search Topic Information:  
· Title of Invention: DOCUMENT MANAGEMENT APPARATUS, SYSTEM AND METHOD

· Inventors: MAXHAM, Jason G.; JENKS, Andrew M.; WORK, Matthew K.; BRENNAN, J. Leslie

· Assignee: DISCOVERY MINING, INC

\*\*\*\*\*

· Background: (CON pages 1-2)  
· Concept of invention: (CON page 3)  
· Claims:  
· 29 (\*29-4) (CLM pages 1-2); Support (CLM-Sup pages 1-2)  
· 34 (\*34-2) (CLM pages 2-3); Support (CLM-Sup pages 2-3)  
· Search hint: "digital signature", "MD5 checksum", "fingerprint", loading

RECEIVED  
NOV 21 2005

BY: -----

substantially equivalent number of documents (converted to same format) to each computer node in the network.

References used before: Burrows (US 5745900A, teaches creating a fingerprint for the entire content of each page), Getchius et al. (US 6493721B1), Narendran et al. (US 6070191A)

Special Instructions and Other Comments:

\* Preferred searcher: ~~Geoffrey St Leger~~

The above referenced pages will be separately submitted to EIC by the examiner to match with this search request.



13281 U.S.PTO  
010704

53217-013

DOCUMENT MANAGEMENT APPARATUS,  
SYSTEM AND METHOD

Related Applications

[01] This application claims priority from Provisional Application Serial No. 60/438,508 filed on January 8, 2003, entitled: "ELECTRONIC DOCUMENT MANAGEMENT", the entire disclosure of which is hereby incorporated by reference herein.

Field of the Invention

[02] The invention relates to an apparatus and method for storing, searching, retrieving, and delivering electronic documents, and a program product for implementing the same, for the purpose of managing a multiplicity of documents.

Background

[03] Many of today's businesses employ sophisticated document management systems for managing existing electronic documents. Despite this, there has not been developed a document management system for providing management services to both existing electronic documents and paper documents. Of particular importance is the need to provide an effective search tool for documents, for example, produced during litigation. Current products on the market permit users to scan paper-based documents or convert electronic documents to a standard format, such as TIFF. However, conversion of tremendous amounts of documents can be time consuming, and expensive. Moreover, document conversion does not reliably maintain all information in a respective document across the many types of file types that may be examined.

[04] Also, in court litigation and regulatory proceedings, prior electronic document management structures and methods to store, search, retrieve and deliver electronic

documents generally require a constrained format to accomplish the necessary functions to achieve effective electronic document management. This is particularly the case in litigation matters where a party before a court needs to organize a multiplicity of documents into a manageable electronic document system. In such litigation, the documents take a variety of formats and structures ranging from letters to detailed reports so that a rigid format may not provide the accessibility and precise recall of critical information for litigation.

[05] A document management system is needed to alleviate the above mentioned problems.

### Summary

[06] The concepts disclosed herein alleviate the above noted problems.

[07] More particularly, a method for managing a plurality of native documents to be uploaded to a document management computer system, includes determining a file type for each native document of the plurality of native documents, creating a fingerprint for each native document, de-duplicating each native document in accordance with the fingerprint, extracting data from each native document, associating extracted data with a corresponding native document, and distributing the plurality of native documents and extracted data substantially equally amongst a plurality of nodes of the document management computer system. By distributing native documents and extracted data substantially equally amongst the nodes, search processing time may be reduced.

[08] Another novel aspect includes a method for searching a plurality of native documents stored in a document management computer system having a plurality of computer nodes storing the plurality of native documents. The steps include defining search criteria for searching the plurality of native documents, executing in parallel searches in accordance with the search criteria for each computer cluster of the plurality of nodes, wherein each computer cluster scores each search result in accordance with the search criteria, ranking the search results in accordance with the score determined in each computer cluster, omitting certain documents represented by the search results in accordance with a user's predefined

Abstract of the Disclosure*Concept*  
7

The concepts herein address a novel document management system including a cluster computer arrangement in which native documents may be stored substantially equally amongst each node. Also disclosed are methods for performing a search based on user-defined search criteria, as well as user-defined classifications that may be applied to documents represented by the search results.

J

- (d) based on results of step (c), receiving search results from the computer system corresponding to the user-defined search query;
- (e) attributing at least one user defined classification to at least one document represented by the search results received in step (d), wherein the user defined classification is displayed when the at least one document is later viewed.

28. (Withdrawn) A method for producing search results of a plurality of native documents stored in a computer system in accordance with a user-defined search query comprising:

- a) providing a Website hosted by a server interfacing with the computer system and a user connected via a user interface over a communication network;
- b) under control of the user interface, displaying the search results of the plurality of native documents in accordance with the user-defined search query; and
- c) in response to at least one user-defined classification selected by the user, attributing the user-defined classification to at least one native document represented by the search results, wherein the user-defined attribute is displayed when the link representing the at least one native document is later viewed.

29. (Currently Amended) An electronic document management system comprising:

29-1 a plurality of computer nodes for storing a plurality of native documents; and

29-2 a computer in communication with the plurality of computer nodes for receiving a

plurality of input files to be uploaded to the plurality of computer nodes,

29-3-1

29-3 wherein the computer is configured to determine the type of native document for each of

29-3-1-1

29-3-1-2 the plurality of input files, to assign a unique identification tag to each native document, and to

29-3-2

eliminate duplicate native documents based on the unique identification tags, for producing a subset of input files to be uploaded to the plurality of computer nodes,

\* 29-4 29-4-1 wherein the subset of input files are distributed substantially equally amongst the

29-4-2 plurality of computer nodes after the identification tags have been assigned and the duplicates have been eliminated.

L 30. (Original) The electronic document management system according to claim 29, wherein the computer is further configured to extract data from each native document.

31. (Original) An electronic document management system according to claim 30, wherein the computer creates a text file corresponding to the extracted data.

32. (Original) An electronic document management system according to claim 29, wherein the computer creates a data file selected from a group consisting of a text file, a meta data file, a XML file, and a HTML file.

33. (Original) An electronic document management system according to claim 29, wherein the subset of input files and associated data extracted therefrom are distributed substantially equally amongst the plurality of computer nodes.

34. (Currently Amended) An electronic document management system comprising a  
34-1-2 PC type computer connected in a parallel cluster, said computer using an operating system that  
34-2 stores electronic documents substantially equally in number throughout the cluster, said  
34-3 operating system generating a fingerprint for each document, where each document is identified  
34-4 by its file extension, converted to ASCII text and given a unique identification number, each of a  
34-5 34-6 34-7 plurality of documents having at least one of either meta-data, text or attachments that are  
34-8

*34-9*

indexed for web-based retrieval from the cluster, said plurality of documents are de-duplicated in  
*34-10*  
accordance with its fingerprint, said plurality of documents forming a cluster data base that is  
*34-11*  
web-searchable by use of a predetermined descriptive term.

infrared (IR) data communications. Common forms of computer-readable media therefore include, for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD, any other optical medium, less commonly used media such as punch cards, paper tape, any other physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer can read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

[33] Referring again to Fig. 1, the computer system 10 may be accessible by an administrator via a stand-alone work station represented by computer 12. An internet server 20 interfaces with the computer system 10 to permit end-user access the system via the internet 24 through at least one user terminal 22.

[34] The computer system 10 is configured to manage large sets of documents for multiple clients, but limits user access to documents supplied by the associated client. Documents supplied by a client are uploaded to the computer system 10 using work station 12. Documents may be supplied in electronic form or in hard copy form. If in electronic form, a suitable drive 26 corresponding to the medium type is used to upload electronic documents to the computer system 10. Also, if documents are in hard copy, they may be scanned using scanner 28 and uploaded to the computer system 10.

[35] Fig. 2 illustrates a system for managing documents uploaded to the computer system 10. First, data is loaded into the computer system 10 via workstation 12. Next, the file type discriminator 212 determines file types based on the file extension of each input file 210. If the file type is an archive, such as .zip, .tar, etc., archive extractor 214 extracts archived file(s). Again, the file type of the extracted documents are determined by the file type discriminator 212.

✓  
cl-29

[36] Often clients periodically upload documents to the computer system 10 and provide large document sets to be uploaded at any one time. As a result, duplicate documents may be

stored in the computer system 10. Also, duplicate documents may exist amongst the documents to be uploaded. Before distributing input files 210 in the computer system 10, file categorizer 216 creates a fingerprint of each file. Well known cryptographic algorithms, such as the MD5 checksum, may be used to create a fingerprint unique to each file. In accordance with the fingerprint, each document is de-duplicated. More particularly, de-duplicator 218 compares the fingerprint of each input file 210 with other fingerprints corresponding to the other input files 210, and compares with the fingerprints of documents already stored in the computer system 10. If a match is found, the document to be uploaded is discarded, so as to prevent multiple documents from residing in the computer system 10.

[37] After the documents to be uploaded have been de-duplicated, extractor 220 converts each native document 222 (corresponding to the input files 210 in original format) to at least a text file 224. Other files that may be generated include meta data files 226, XML files 228, and HTML files 230. Well known third party software packages may be used in this conversion process.

② 34 ↓ [38] Indexer 232 creates a file association table for each native document that maintains the associations between each native document 222, converted documents 224-230, and attachments, if any, to the native document. These attachments commonly referred to as "children files." While the file association table may be stored in any of the nodes N1-Nn, other databases 16a-c may be used to maintain file association tables. Distributor 234 distributes native documents and converted documents substantially equally amongst the nodes of the computer system 10, after which time, the documents may be searched. ② 29 ↑

[39] Referring back to Fig. 1, a three cluster arrangement is shown. In this example, about a third of the documents to be uploaded would be distributed to each node N1 - Nn of the computer system 10. Each processor 18 interfacing within each node Nn executes a search daemon for searching files in each node. Therefore, when a search is initiated by server 20, multiple processors 18 execute the search in parallel. The search daemon scores each document based on search criteria specified. Results from each search daemon can be

compared against results from other search daemons. For example, Table 1 provides an example of search results produced by each search daemon.

Table 1

*(d 34  
1*

Search Results	Node N1 Results Scoring	Node N2 Results Scoring	Node N3 Results Scoring
Document 1	0.96	0.99	0.95
Document 2	0.82	0.80	0.93
Document 3	0.76	0.45	0.77
Document 4	0.50	N/A	0.39
Document 5	0.49	N/A	0.25

[40] Server 20 receives search results from each processor 18 and merges the search results accordingly. Assuming that only the top five search results were requested, the search results may be compiled in the following manner.

Table 2

Main Results	Location	Score
1.	Document 1, Node N2	0.99
2.	Document 1, Node N1	0.96
3.	Document 1, Node N3	0.95
4.	Document 2, Node N3	0.93
5.	Document 2, Node N1	0.82

[41] In more detail, Fig. 3 illustrates a flow diagram of the search process initiated by server 20. First, in Step 310, server 20 receives a search query from a user via a user interface 22 over the internet 24. In Step 312, server 20 initiates the parallel query tool, *i.e.*, server 20 causes each processor 18 to execute respective search queries in accordance with the search criteria received by server 20. In Step 314, server 20 receives the search results from each processor 18 of each cluster, *e.g.*, as shown in Table 1.

[42] Users accessing the computer system may have pre-defined permission levels, *e.g.*, on a scale of 1 to 5; 1 being the lowest level and 5 being the highest. Also, documents classifications may be assigned to each document on the same scale. Therefore, only documents that have a document classification equal to or less than the user's pre-defined

Set	Items	Description
S1	3	AU=((MAXHAM J? OR MAXHAM, J?) AND (JENKS A? OR JENKS, A?) - AND (WORK M? OR WORK, M?) AND (BRENNAN, J? OR BRENNAN J?))
S2	111	AU=(MAXHAM J? OR MAXHAM, J? OR JENKS A? OR JENKS, A? OR WORK M? OR WORK, M? OR BRENNAN, J? OR BRENNAN J?) AND (DOCUMENT-????? OR FILE? ? OR HYPERTEXT? ? OR REPORT? ? OR LETTER? ? OR EMAIL??? OR E()MAIL??? ) AND AY=(1963:2001)
S3	74	S2 AND (NODE? ? OR SUPERNODE? ? OR ADDRESS? ? OR SERVER? ? OR GATEWAY? ? OR CPU OR CENTRAL()PROCESSING()UNIT OR COMPUTER? ? OR CLIENT? ? OR WORKSTATION? ? OR MICROCOMPUTER? ? OR MICROPROCESSOR? ? OR MICROPROCESSER? ? OR PC? ? OR MASTER? ? OR SLAVE? ?)
S4	44	S3 AND (ID? ? OR TAG? ? OR IDENTIFY??? OR IDENTIFIC????? OR IDENTIFIER? ? OR FLAG? ? OR NAME? ? OR NUMBER? ? OR INDICATOR? ? OR FINGERPRINT? ? OR MD5()CHECKSUM)

? show files

File 347:JAPIO Nov 1976-2005/Jul (Updated 051102)  
(c) 2005 JPO & JAPIO

File 348:EUROPEAN PATENTS 1978-2005/Dec W02  
(c) 2005 European Patent Office

File 349:PCT FULLTEXT 1979-2005/UB=20051215,UT=20051208  
(c) 2005 WIPO/Univentio

File 350:Derwent WPIX 1963-2005/UD,UM &UP=200581  
(c) 2005 Thomson Derwent

*patent inventor search*

1/5/3 (Item 1 from file: 350)  
X DIALOG(R) File 350:Derwent WPIX  
(c) 2005 Thomson Derwent. All rts. reserv.

my application

016395920 \*\*Image available\*\*  
WPI Acc No: 2004-553829/200453  
XRPX Acc No: N04-438194

Native documents managing method for business application, involves distributing set of native documents and extracted data, substantially amongst set of nodes on document management computer system

Patent Assignee: BRENNAN J L (BREN-I); JENKS A M (JENK-I); MAXHAM J G (MAXH-I); WORK M K (WORK-I); DISCOVERY MINING INC (DISC-N)

Inventor: BRENNAN J L ; JENKS A M ; MAXHAM J G ; WORK M K

Number of Countries: 108 Number of Patents: 002

Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week
WO 200463863	A2	20040729	WO 2004US168	A	20040107	200453 B
US 20040187075	A1	20040923	US 2003438508	P	20030108	200463
			US 2004752432	A	20040107	

Priority Applications (No Type Date): US 2003438508 P 20030108; US 2004752432 A 20040107

Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
WO 200463863	A2	E	38	G06F-000/00	

Designated States (National): AE AG AL AM AT AU AZ BA BB BG BR BW BY BZ CA CH CN CO CR CU CZ DE DK DM DZ EC EE EG ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA MD MG MK MN MW MX MZ NA NI NO NZ OM PG PH PL PT RO RU SC SD SE SG SK SL SY TJ TM TN TR TT TZ UA UG US UZ VC VN YU ZA ZM ZW

Designated States (Regional): AT BE BG BW CH CY CZ DE DK EA EE ES FI FR GB GH GM GR HU IE IT KE LS LU MC MW MZ NL OA PT RO SD SE SI SK SL SZ TR TZ UG ZM ZW

US 20040187075 A1 G06F-017/00 Provisional application US 2003438508

Abstract (Basic): WO 200463863 A2

NOVELTY - The method involves creating a fingerprint for each native document with a file type. Each native document is de-duplicated in accordance with a created fingerprint. A data is extracted from each document and extracted data is associated with the corresponding native document. The set of documents and extracted data are distributed substantially amongst a set of nodes on a document management computer system.

DETAILED DESCRIPTION - INDEPENDENT CLAIMS are also included for the following:

- a program product for managing native documents
- a method of searching a set of native documents stored in a document management computer system
- a program product for searching a set of native documents stored in a document management computer system
- an electronic document management system.

USE - Used for managing native document to be uploaded to a document management computer system in business application.

ADVANTAGE - The set of documents and extracted data are distributed substantially amongst a set of nodes, thereby reducing the search processing time.

DESCRIPTION OF DRAWING(S) - The drawing shows a computer system used to implement the native documents managing method.

Computers (12, 22)  
Disk drives (14)  
Processor (18)

Server (20)  
pp; 38 DwgNo 1/11  
Title Terms: NATIVE; DOCUMENT; MANAGE; METHOD; BUSINESS; APPLY; DISTRIBUTE;  
SET; NATIVE; DOCUMENT; EXTRACT; DATA; SUBSTANTIAL; SET; NODE; DOCUMENT;  
MANAGEMENT; COMPUTER; SYSTEM  
Derwent Class: T01  
International Patent Class (Main): G06F-000/00; G06F-017/00  
File Segment: EPI  
?

Set	Items	Description
S1	0	AU=((MAXHAM J? OR MAXHAM, J?) AND (JENKS A? OR JENKS, A?) - AND (WORK M? OR WORK, M?) AND (BRENNAN, J? OR BRENNAN J?))
S2	136	AU=(MAXHAM J? OR MAXHAM, J? OR JENKS A? OR JENKS, A? OR WORK M? OR WORK, M? OR BRENNAN, J? OR BRENNAN J?) AND (DOCUMENT-????? OR FILE? ? OR HYPERTEXT? ? OR REPORT? ? OR LETTER? ? OR EMAIL??? OR E()MAIL??? ) AND (PY<2002 OR PD<20020108)
S3	15	S2 AND (NODE? ? OR SUPERNODE? ? OR ADDRESS? ? OR SERVER? ? OR GATEWAY? ? OR CPU OR CENTRAL()PROCESSING()UNIT OR COMPUTER? ? OR CLIENT? ? OR WORKSTATION? ? OR MICROCOMPUTER? ? OR MICROPROCESSOR? ? OR MICROPROCESSER? ? OR PC? ? OR MASTER? ? OR SLAVE? ?)
? show files		
File	2:INSPEC 1898-2005/Dec W2	(c) 2005 Institution of Electrical Engineers
File	6:NTIS 1964-2005/Dec W2	(c) 2005 NTIS, Intl Cpyrght All Rights Res
File	8:Ei Compendex(R) 1970-2005/Dec W2	(c) 2005 Elsevier Eng. Info. Inc.
File	34:SciSearch(R) Cited Ref Sci 1990-2005/Dec W2	(c) 2005 Inst for Sci Info
File	65:Inside Conferences 1993-2005/Dec W3	(c) 2005 BLDSC all rts. reserv.
File	94:JICST-EPlus 1985-2005/Oct W3	(c) 2005 Japan Science and Tech Corp(JST)
File	99:Wilson Appl. Sci & Tech Abs 1983-2005/Oct	(c) 2005 The HW Wilson Co.
File	148:Gale Group Trade & Industry DB 1976-2005/Dec 19	(c) 2005 The Gale Group
File	636:Gale Group Newsletter DB(TM) 1987-2005/Dec 19	(c) 2005 The Gale Group
?		

*NPL inventor*

Set	Items	Description
S1	1981159	NODE? ? OR SUPERNODE? ? OR ADDRESS? ? OR SERVER? ? OR GATEWAY? ? OR CPU OR CENTRAL()PROCESSING()UNIT OR COMPUTER? ? OR - CLIENT? ? OR WORKSTATION? ? OR MICROCOMPUTER? ? OR MICROPROCESSOR? ? OR MICROPROCESER? ? OR PC? ? OR MASTER? ? OR SLAVE? ?
S2	339972	DOCUMENT????? OR FILE? ? OR HYPERTEXT? ? OR REPORT? ? OR LETTER? ? OR EMAIL??? OR E()MAIL???
S3	702808	DISTRIBUT??? OR DISSEMINAT??? OR DISEMINAT???
S4	3961999	ADD OR ADDING OR ADDED OR ASSIGN???? OR ASSOCIAT??? OR GENERAT???
S5	5950338	UNIQUE OR ONE OR INDIVIDUAL OR SINGLE OR SINGULAR??? OR SOLITARY OR SPECIFIC???? OR EXCLUD??? OR EXCLUS???
S6	2886621	ELIMINAT??? OR REMOV??? OR DELET??? OR DISCARD??? OR EXTRACT??? OR (TAKE OR TAKING OR TOOK OR CUT) ()OUT OR ERAS???
S7	10239	(ID? ? OR TAG? ? OR IDENTIFY??? OR IDENTIFIC????? OR IDENTIFIER? ? OR FLAG? ? OR NAME? ? OR NUMBER? ? OR INDICATOR? ? OR FINGERPRINT? ? OR MD5()CHECKSUM OR MARKER? ? OR LABEL? ?) (3N- ) S3
S8	28479	S6 (3N) (DUPLICATE? ? OR REDUNDAN???? OR COPY? ? OR RECURR??- ?? OR REPLICAT??? OR SECOND OR TWIN OR REPLICA? ?)
S9	35515	(EQUAL?? OR EQUITAB?? OR PROPORTIONA???? OR EQUIVALENT?? OR EVEN?? OR IDENTICAL?? OR LIKEWISE OR SIMILAR?? OR UNIFORM?? - OR ALIKE OR "SAME") (3N) S1
S10	1355	S4 AND S5 AND S7
S11	6	S10 AND S8
S12	0	S11 AND S2 AND S3 AND S9
S13	1	(S10 AND S2 AND S3 AND S9) NOT S11
S14	271	(S2 AND S3 AND S9) NOT (S11 OR S13)
S15	250	S14 AND (MC=T01 OR IC=(G06F) OR IC=(H04L))
S16	120	S14 AND (MC=T01 OR IC=(G06F-017?) OR IC=(G06F-007?) OR IC==(G06F-000?) OR IC=(H04L))
S17	53	S16 AND AY=(1963:2001)
? show files		
File 347:JAPIO Nov 1976-2005/Jul (Updated 051102)		
(c) 2005 JPO & JAPIO		
File 350:Derwent WPIX 1963-2005/UD,UM &UP=200581		
(c) 2005 Thomson Derwent		
?		

Bibliographic patent

Full Text Patent

Set	Items	Description
S1	1722365	NODE? ? OR SUPERNODE? ? OR ADDRESS? ? OR SERVER? ? OR GATEWAY? ? OR CPU OR CENTRAL() PROCESSING() UNIT OR COMPUTER? ? OR - CLIENT? ? OR WORKSTATION? ? OR MICROCOMPUTER? ? OR MICROPROCESSOR? ? OR MICROPROCESSER? ? OR PC? ? OR MASTER? ? OR SLAVE? ?
S2	1058304	STOR??? OR SAVE OR SAVING OR SAVED OR RECORD??? OR ACCUMULATE??? OR KEEP??? OR CACH???
S3	2255772	DOCUMENT????? OR FILE? ? OR HYPERTEXT? ? OR REPORT? ? OR LETTER? ? OR EMAIL??? OR E()MAIL???
S4	1614902	UPLOAD? ? OR RECEIV??? OR RECEPCT??? OR RETRIEV??? OR FETCH??? OR ACQUIR??? OR OBTAIN??? OR DOWNLOAD??? OR RECEIPT OR ACCEPT??? OR GET OR GETTING OR GOT OR OBTAIN??? OR DISTRIBUT???
S5	1686899	DISPATCH??? OR DOWNLOAD??? OR FORWARD??? OR ROUTE OR ROUTING OR ROUTED OR SEND OR SENDING OR SENT OR TRANSFER??? OR TRANSMIS??? OR TRANSMIT??? OR DISTRIBUT?? OR DELIVER??? OR REMIT???? OR SHIP???? OR DISSEMINAT??? OR DISEMINAT???
S6	1560814	ADD OR ADDING OR ADDED OR ASSIGN??? OR ASSOCIAT??? OR GENERAT???
S7	1924912	UNIQUE OR ONE OR INDIVIDUAL OR SINGLE OR SINGULAR??? OR SOLITARY OR SPECIFIC???? OR EXCLUD??? OR EXCLUS???
S8	1204410	ELIMINAT??? OR REMOV??? OR DELET??? OR DISCARD??? OR EXTRACT??? OR (TAKE OR TAKING OR TOOK OR CUT) () OUT OR ERAS???
S9	82558	(ID? ? OR TAG? ? OR IDENTIFY??? OR IDENTIFIC????? OR IDENTIFIER? ? OR FLAG? ? OR NAME? ? OR NUMBER? ? OR INDICATOR? ? OR FINGERPRINT? ? OR MD5() CHECKSUM OR MARKER? ? OR LABEL? ?) (3N- ) S3
S10	57816	S8 (3N) (DUPLICATE? ? OR REDUNDAN??? OR COPY? ? OR RECURR??- ?? OR REPLICAT??? OR SECOND OR TWIN OR REPLICA? ?)
S11	103725	(EQUAL?? OR EQUITAB?? OR PROPORTIONA???? OR EQUIVALENT?? OR EVEN?? OR IDENTICAL?? OR LIKEWISE OR SIMILAR?? OR UNIFORM?? - OR ALIKE OR "SAME") (3N) S1
S12	1488	S6 (3N) S7 (3N) S9
S13	41	S10 (3N) S9
S14	1532	S3 (3N) (S2 OR S4 OR S5) (3N) S11
S15	0	S12 (100N) S13 (100N) S14
S16	0	S13 (100N) S14
S17	13	S12 (100N) S14
S18	266	(S14 AND (IC=(G06F-017/30) OR IC=(G06F-007/00) OR IC=G06F-- 017/00)) NOT S17
S19	20	(S6 (3N) S7 (3N) S14) NOT S17
S20	3	(S10 (3N) S14) NOT (S17 OR S19)
S21	198	(S18 AND AY=(1963:2001)) NOT (S17 OR S19:S20)
S22	2	S21 AND S11/TI
S23	30	(S21 AND (S3 AND (S2 OR S4 OR S5)) /TI) NOT (S17 OR S19:S20 OR S22)
S24	521145	DISTRIBUT?? OR DISSEMINAT?? OR DISEMINAT???
S25	90	(S3 (3N) S24 (3N) S11) NOT (S17 OR S19:S20 OR S22:S23)
S26	78	S25 AND AY=(1963:2001)
S27	0	S26 AND S11/TI
S28	5	S26 AND (S3 AND (S2 OR S4 OR S5)) /TI
S29	73	S26 NOT S28

? show files

File 348:EUROPEAN PATENTS 1978-2005/Dec W02

(c) 2005 European Patent Office

File 349:PCT FULLTEXT 1979-2005/UB=20051215,UT=20051208

(c) 2005 WIPO/Univentio

?

29/3,K/49 (Item 26 from file: 349)

DIALOG(R) File 349:PCT FULLTEXT

(c) 2005 WIPO/Univentio. All rts. reserv.

00828024 \*\*Image available\*\*

**HASH FILE SYSTEM AND METHOD FOR USE IN A COMMONALITY FACTORING SYSTEM  
SYSTEME ET PROCEDE DE FICHIERS DE HACHAGE UTILES DANS UN SYSTEME DE  
FACTORIZATION D'ELEMENTS COMMUNS**

Patent Applicant/Assignee:

UNDOO INC, 1 Technology Drive, Building A, Irvine, CA 92618, US, US  
(Residence), US (Nationality), (For all designated states except: US)

Patent Applicant/Inventor:

MOULTON Gregory Hagan, 6 Bayberry Way, Irvine, CA 92612, US, US  
(Residence), US (Nationality), (Designated only for: US)  
WHITEHILL Stephen B, 12880 Crawford Drive, Tustin, CA 92782, US, US  
(Residence), US (Nationality), (Designated only for: US)

Legal Representative:

BURTON Carol W (et al) (agent), Hogan & Hartson LLP, Suite 1500, 1200  
17th Street, Denver, CO 80202, US,

Patent and Priority Information (Country, Number, Date):

Patent: WO 200161563 A1 20010823 (WO 0161563)

Application: WO 2001US4763 20010214 (PCT/WO US0104763)

Priority Application: US 2000183762 20000218; US 2000245920 20001106; US  
2001777150 20010205

Designated States:

(Protection type is "patent" unless otherwise stated - for applications  
prior to 2004)

AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE  
ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT  
LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM  
TR TT TZ UA UG US UZ VN YU ZA ZW  
(EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR  
(OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG  
(AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW  
(EA) AM AZ BY KG KZ MD RU TJ TM

Publication Language: English

Filing Language: English

Fulltext Word Count: 11043

Fulltext Availability:

Detailed Description

Detailed Description

... to efficiently distribute web content because the  
method of factoring commonality (hashing) also produces uniform  
distribution over all hash file system servers. This even  
distribution  
permits a large array of servers to function as a gigantic web server  
2 0...

29/3, K/29 (Item 6 from file: 349)

DIALOG(R) File 349:PCT FULLTEXT

(c) 2005 WIPO/Univentio. All rts. reserv.

00857665 \*\*Image available\*\*

**SECURED ELECTRONIC MAIL SYSTEM AND METHOD**  
**SYSTEME ET PROCEDE DE COURRIER ELECTRONIQUE SECURISE**

Patent Applicant/Assignee:

V N HERMES INC, 32 North Dean Street, Englewood, NJ 07631-2807, US, US  
(Residence), US (Nationality)

Inventor(s):

NEMOVICHER C Kerry, 39 Markham Circle, Englewood, NJ 07631, US,

Legal Representative:

SCHEER Michael J (et al) (agent), Ostrolenk, Faber, Gerb & Soffen, LLP,  
1180 Avenue of the Americas, New York, NY 10036, US,

Patent and Priority Information (Country, Number, Date):

Patent: WO 200191403 A2-A3 20011129 (WO 0191403)

Application: WO 2001US16714 20010523 (PCT/WO US0116714)

Priority Application: US 2000206580 20000523

Designated States:

(Protection type is "patent" unless otherwise stated - for applications prior to 2004)

AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE  
ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT  
LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM  
TR TT TZ UA UG UZ VN YU ZA ZW  
(EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR  
(OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG  
(AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW  
(EA) AM AZ BY KG KZ MD RU TJ TM

Publication Language: English

Filing Language: English

Fulltext Word Count: 16660

Fulltext Availability:

Detailed Description

Detailed Description

... become inoperable, primary node 28 begins receiving all e-mail from communication network 130, and **distributes** the e-mail to all other nodes in an even distribution or load balancing process. That (inverted exclamation mark)S, primary node 28 takes over the...

# Bibliographic NPL

Set	Items	Description
S1	7780458	NODE? ? OR SUPERNODE? ? OR ADDRESS? ? OR SERVER? ? OR GATEWAY? ? OR CPU OR CENTRAL()PROCESSING()UNIT OR COMPUTER? ? OR -CLIENT? ? OR WORKSTATION? ? OR MICROCOMPUTER? ? OR MICROPROCESSOR? ? OR MICROPROCESSEER? ? OR PC? ? OR MASTER? ? OR SLAVE? ?
S2	6065593	DOCUMENT????? OR FILE? ? OR HYPertext? ? OR REPORT? ? OR LETTER? ? OR EMAIL??? OR E()MAIL???
S3	4837678	DISTRIBUT??? OR DISSEMINAT??? OR DISEMINAT???
S4	9322919	ADD OR ADDING OR ADDED OR ASSIGN??? OR ASSOCIAT??? OR GENERAT???
S5	14626311	UNIQUE OR ONE OR INDIVIDUAL OR SINGLE OR SINGULAR??? OR SOLITARY OR SPECIFIC???? OR EXCLUD??? OR EXCLUS???
S6	3500440	ELIMINAT??? OR REMOV??? OR DELET??? OR DISCARD??? OR EXTRACT??? OR (TAKE OR TAKING OR TOOK OR CUT)()OUT OR ERAS???
S7	63318	(ID? ? OR TAG? ? OR IDENTIFY??? OR IDENTIFIC????? OR IDENTIFIER? ? OR FLAG? ? OR NAME? ? OR NUMBER? ? OR INDICATOR? ? OR FINGERPRINT? ? OR MD5()CHECKSUM OR MARKER? ? OR LABEL? ?) (3N-)S3
S8	32887	S6(3N) (DUPLICATE? ? OR REDUNDAN???? OR COPY? ? OR RECURR??-? ? OR REPLICAT??? OR SECOND OR TWIN OR REPLICA? ?)
S9	82679	(EQUAL?? OR EQUITAB?? OR PROPORTIONA???? OR EQUIVALENT?? OR EVEN?? OR IDENTICAL?? OR LIKEWISE OR SIMILAR?? OR UNIFORM?? - OR ALIKE OR "SAME") (3N)S1
S10	7023	S4 AND S5 AND S7
S11	24	S10 AND S8
S12	1	S11 AND S2 AND S3 AND S9
S13	67	(S10 AND S2 AND S3 AND S9) NOT S12
S14	5143	(S2 AND S3 AND S9) NOT S11:S13
S15	27	S2(3N)S3(3N)S9
S16	125	(S2(10N)S3(10N)S9) NOT (S11:S13 AND S15)
S17	92	S16 AND (PY<2002 OR PD<20020108)
? show files		
File	2:INSPEC 1898-2005/Dec W2	(c) 2005 Institution of Electrical Engineers
File	6:NTIS 1964-2005/Dec W2	(c) 2005 NTIS, Intl Cpyrht All Rights Res
File	8:Ei Compendex(R) 1970-2005/Dec W2	(c) 2005 Elsevier Eng. Info. Inc.
File	34:SciSearch(R) Cited Ref Sci 1990-2005/Dec W2	(c) 2005 Inst for Sci Info
File	35:Dissertation Abs Online 1861-2005/Nov	(c) 2005 ProQuest Info&Learning
File	56:Computer and Information Systems Abstracts 1966-2005/Dec	(c) 2005 CSA.
File	57:Electronics & Communications Abstracts 1966-2005/Dec	(c) 2005 CSA.
File	60:ANTE: Abstracts in New Tech & ENGINEER 1966-2005/NOV	(c) 2005 CSA.
File	65:Inside Conferences 1993-2005/Dec W3	(c) 2005 BLDSC all rts. reserv.
File	94:JICST-EPlus 1985-2005/Oct W3	(c) 2005 Japan Science and Tech Corp(JST)
File	95:TEME-Technology & Management 1989-2005/Nov W2	(c) 2005 FIZ TECHNIK
File	99:Wilson Appl. Sci & Tech Abs 1983-2005/Oct	(c) 2005 The HW Wilson Co.
File	111:TGG Natl.Newspaper Index(SM) 1979-2005/Dec 16	(c) 2005 The Gale Group
File	144:Pascal 1973-2005/Dec W2	(c) 2005 INIST/CNRS
File	434:SciSearch(R) Cited Ref Sci 1974-1989/Dec	

(c) 1998 Inst for Sci Info  
File 636:Gale Group Newsletter DB(TM) 1987-2005/Dec 19  
(c) 2005 The Gale Group  
?

# Full Text NPL

Set	Items	Description
S1	16556115	NODE? ? OR SUPERNODE? ? OR ADDRESS? ? OR SERVER? ? OR GATEWAY? ? OR CPU OR CENTRAL()PROCESSING()UNIT OR COMPUTER? ? OR -CLIENT? ? OR WORKSTATION? ? OR MICROCOMPUTER? ? OR MICROPROCESSOR? ? OR MICROPROCESSER? ? OR PC? ? OR MASTER? ? OR SLAVE? ?
S2	17789908	STOR??? OR SAVE OR SAVING OR SAVED OR RECORD??? OR ACCUMULAT??? OR KEEP??? OR CACH???
S3	16315870	DOCUMENT????? OR FILE? ? OR HYPERTEXT? ? OR REPORT? ? OR LETTER? ? OR EMAIL??? OR E()MAIL???
S4	8050624	DISTRIBUT??? OR DISSEMINAT??? OR DISSEMINAT???
S5	5074340	ELIMINAT??? OR REMOV??? OR DELET??? OR DISCARD??? OR EXTRACT??? OR (TAKE OR TAKING OR TOOK OR CUT) ()OUT OR ERAS???
S6	499136	(ID? ? OR TAG? ? OR IDENTIFY??? OR IDENTIFIC????? OR IDENTIFIER? ? OR FLAG? ? OR NAME? ? OR NUMBER? ? OR INDICATOR? ? OR FINGERPRINT? ? OR MD5()CHECKSUM OR MARKER? ? OR LABEL? ?) (3N-)S3
S7	89538	S5(3N) (DUPLICATE? ? OR REDUNDAN????? OR COPY? ? OR RECURR??-? ? OR REPLICAT??? OR SECOND OR TWIN OR REPLICA? ?)
S8	488715	(EQUAL?? OR EQUITAB?? OR PROPORTIONA????? OR EQUIVALENT?? OR EVEN?? OR IDENTICAL?? OR LIKEWISE OR SIMILAR?? OR UNIFORM?? - OR ALIKE OR "SAME") (3N)S1
S9	1174	(ADD OR ADDING OR ADDED OR ASSIGN??? OR ASSOCIAT??? OR GENERAT???) (3N) (UNIQUE OR ONE OR INDIVIDUAL OR SINGLE OR SINGULAR??? OR SOLITARY OR SPECIFIC??? OR EXCLUD??? OR EXCLUS???) (-3N)S6
S10	185	S7(3N)S6
S11	219	S3(3N)S4(3N)S8
S12	0	S9(100N)S10(100N)S11
S13	0	S9(100N)S11
S14	0	S10(100N)S11
S15	0	(ADD OR ADDING OR ADDED OR ASSIGN??? OR ASSOCIAT??? OR GENERAT???) (3N) (UNIQUE OR ONE OR INDIVIDUAL OR SINGLE OR SINGULAR??? OR SOLITARY OR SPECIFIC??? OR EXCLUD??? OR EXCLUS???) (-3N)S11
S16	0	S7(3N)S11
S17	187	S11 AND (PY<2002 OR PD<20020108)
S18	1	S17 AND (EQUAL?? OR EQUITAB?? OR PROPORTIONA????? OR EQUIVALENT?? OR EVEN?? OR IDENTICAL?? OR LIKEWISE OR SIMILAR?? OR UNIFORM?? OR ALIKE OR "SAME")/TI
S19	6	(S17 AND (S3 AND S4)/TI) NOT S18
S20	10	(S17 AND S3/TI) NOT (S18 OR S19)
S21	20	(S17 AND S4/TI) NOT (S18:S20)
S22	860	(S3(10N)S4(10N)S8) NOT S17
S23	0	S9(100N)S10(100N)S22
S24	0	S9(100N)S22
S25	0	S10(100N)S22
S26	12	(ADD OR ADDING OR ADDED OR ASSIGN??? OR ASSOCIAT??? OR GENERAT???) (3N) (UNIQUE OR ONE OR INDIVIDUAL OR SINGLE OR SINGULAR??? OR SOLITARY OR SPECIFIC??? OR EXCLUD??? OR EXCLUS???) (-3N)S22
S27	3	S7(3N)S22
S28	698	(S22 AND (PY<2002 OR PD<20020108)) NOT (S26:S27)
S29	24	S28 AND (EQUAL?? OR EQUITAB?? OR PROPORTIONA????? OR EQUIVALENT?? OR EVEN?? OR IDENTICAL?? OR LIKEWISE OR SIMILAR?? OR UNIFORM?? OR ALIKE OR "SAME")/TI
S30	8	(S28 AND (S3 AND S4)/TI) NOT (S26:S27 OR S29)
S31	75	(S28 AND S3/TI) NOT (S26:S27 OR S29:S30)
S32	39	(S28 AND S4/TI) NOT (S26:S27 OR S29:S31)

32/9/20 (Item 1 from file: 88)  
DIALOG(R)File 88:Gale Group Business A.R.T.S.  
(c) 2005 The Gale Group. All rts. reserv.

05442318 SUPPLIER NUMBER: 62589963  
**Evaluating the Performance of Distributed Architectures for Information Retrieval Using a Variety of Workloads. (Statistical Data Included)**  
CAHOON, BRENDON; MCKINLEY, KATHRYN S.; LU, ZHIHONG  
ACM Transactions on Information Systems, 18, 1, 1  
Jan, 2000  
DOCUMENT TYPE: Statistical Data Included ISSN: 1046-8188  
LANGUAGE: English RECORD TYPE: Fulltext  
WORD COUNT: 15598 LINE COUNT: 01344

TEXT:

The information explosion across the Internet and elsewhere offers access to an increasing number of document collections. In order for users to effectively access these collections, information retrieval (IR) systems must provide coordinated, concurrent, and distributed access. In this article, we explore how to achieve scalable performance in a distributed system for collection sizes ranging from 1GB to 128GB. We implement a fully functional distributed IR system based on a multithreaded version of the Inquery unified IR system. To explore the design space more fully, we also implement and validate a flexible simulation model. We measure performance as a function of system parameters such as client command rate, number of document collections, terms per query, query term frequency, number of answers returned, and command mixture. Our results show that it is important to model both query and document commands because the heterogeneity of commands significantly impacts performance. Based on our results, we recommend simple changes to the prototype and evaluate the changes using the simulator. Because of the significant resource demands of information retrieval, it is not difficult to generate workloads that overwhelm system resources regardless of the architecture. However under some realistic workloads, we demonstrate system organizations for which response time gracefully degrades as the workload increases and performance scales with the number of processors. This scalable architecture includes a surprisingly small number of brokers through which a large number of clients and servers communicate.

#### 1. INTRODUCTION

The increasing numbers of large, unstructured text collections require full-text information retrieval (IR) systems in order for users to access them effectively. Current systems typically only allow users to connect to a single database either locally or perhaps on another machine. A distributed IR system should be able to provide multiple users with concurrent, efficient access to multiple text collections located on disparate sites. Since the documents in unstructured text collections are independent, IR systems are ideal applications to distribute across a network of workstations. However, the high resource demands of IR systems limit their performance, especially as the number of users, as well as the size and number of text collections, increases. Distributed computing offers a solution to these problems.

Only recently have people published work on distributed architectures for information retrieval. The Very Large Collection track in the TREC conferences promotes the development of distributed and shared memory architectures for IR (Hawking and Thistlewaite 1997; Hawking et al. 1998). Several researchers created distributed IR systems and demonstrated the feasibility of distributed architectures for information retrieval (Harman et al. 1991; Macleod et al. 1987). However, it is not clear from these initial implementations how the systems will perform in practice, since, unlike the case for database systems, very little experimental investigation of distributed IR systems has been performed, and researchers

have typically limited their collection sizes to at most 1GB. The focus of this article is to design large-scale distributed IR architectures by analyzing the performance of potential systems under a variety of workloads. In our work, we simulate configurations with up to 128 servers, each containing a 1GB text collection (i.e., 128GB of total text). We begin with a prototype implementation of a distributed information retrieval system using Inquery; an inference network, full-text information retrieval model (Callan et al. 1992). Our system adopts a variation of the client-server paradigm that consists of clients connected to information retrieval engines through a central administration broker, as we illustrate in Figure 1.

(Figure 1 ILLUSTRATION OMITTED)

In our experiments, we model a simple hardware configuration using a single CPU and disk. In our prototype, we use a multithreaded version of the Inquery server to obtain high machine utilization. We illustrate that using a single-threaded Inquery server does not effectively utilize CPU and disk resources; we achieve the best performance when we use four threads. We use this configuration for our base system. A symmetric multiprocessor with multiple disks clearly improves performance as well (Lu et al. 1998), but, in this article, we focus on improvements in a system distributed over a network of workstations.

In the original Inquery system (not distributed or multithreaded), clients specify an Inquery text collection, connect to it, interact with it, and finally disconnect. In the prototype distributed system, clients search multiple text collections simultaneously. To build our prototype, we made the fewest possible changes to the underlying software. As illustrated in Figure 1, we use a single central broker, which we call the connection server, to administrate connections between clients and the IR engines. We refer to the IR engines as Inquery servers. The connection server maintains a list of available collections and their locations, and brokers all of the clients' retrieval requests and Inquery server responses. We describe this distributed system in detail in Section 3. We measure the system and use it to drive a simulator in which we can easily move and replicate functionality to investigate alternative architectures for our distributed system. Section 4 presents this simulation model. We validate our model by comparing the performance of the simulator to the actual system for query operations.

We parameterize our simulation model by system features such as the number of CPUs, number of disks, and network utilization. This model allows us to investigate systems that vary from our implementation. In Section 5, we measure system performance including response time and resource utilization for a variety of configurations. During our investigation we identify bottlenecks and study the effects of various architectures and parameters. Our goal is to use resources efficiently by maximizing parallelism and ensuring scalability. We also maintain the effectiveness, in terms of recall and precision (Callan et al. 1995a), of a standalone IR system.

Clients model the activities of realistic users during our experiments. Each client issues query commands, obtains summary information for query results, and retrieves documents. We simulate client think time by varying the client command rate; a slower command rate simulates a larger amount of think time. The different IR commands, i.e., query and document, have distinct effects on performance. For example, we show the architecture is able exploit parallelism in the summary information commands, which improves response times as the size of the architecture increases, up to eight or 32 Inquery servers.

In our main experiments, we explore an architecture that distributes multiple, independent text collections. As the number of machines increases, the amount of text to search also increases. We believe the multiple-text-collection architecture is similar to actual systems, especially for large architectures, and we present detailed experiments.(1)

We demonstrate performance as a factor of the client command rate, number of text collections, the number of terms per query, the query term frequencies, number of answers returned, and the mixture of query, summary, and document retrieval operations. In order to generalize our results, we experiment with a range of values for these parameters and discuss their effects on system performance.

When clients issue short queries, response times are reasonable for many different numbers of clients and Inquiry servers. However, the connection server is a bottleneck when (1) clients issue commands quickly and (2) there are a large number of Inquiry servers. We show that adding a small number of brokers to manage the clients and Inquiry servers is an effective strategy for improving performance and ensuring scalability for short queries. The architecture does not perform well when users enter long queries, due to bottlenecks in the Inquiry servers. We show that adding more hardware, such as CPUs and disks, enables us to add more threads and alleviate this bottleneck.

We, thus, present a large-scale distributed IR architecture that provides efficient access to collection sizes up to 128GB by exploiting the heterogeneity among query and document operations. This article also identifies several important factors that system architects must consider when designing distributed IR systems.

We organize the rest of the article as follows. Section 2 presents the related work. We describe our distributed IR system in Section 3. In Section 4 we describe our simulation model. We list the simulator's timing values, validate the model, and describe the parameters we use in our experiments. Using the simulator, we measure the performance of the distributed IR system in Section 5. We also experiment with several changes in order to improve performance. Finally, we summarize our work in Section 6.

## 2. RELATED WORK

Related work on architectures for distributed IR systems include designing and evaluating architecture performance, data partitioning, caching, and multiprocessor systems. There is also a large volume of related work in database systems, so we only briefly discuss the differences between IR and database systems.

The work on architecture performance is most closely related to this work. Our research combines and extends previous work in distributed IR, since we model and analyze a complete system architecture. Although others have examined some of the issues, no one has considered the entire system under a variety of workloads and conditions. We base our distributed system on a proven, effective retrieval engine. We also model architectures with very large text collections -- up to 128GB of data on 128 Inquiry servers. Much of the prior work on distributed IR architectures has been restricted to small collections, typically less than 1GB and/or 16 servers. Participants in the TREC Very Large Collection track use collections up to 100GB, but they only provide query-processing times for a single query at a time, rather than a variety of workloads as we do here (Hawking et al. 1998). It is clear that some industrial sites have collections even larger than ours, but they choose not to report on them in the literature to maintain their competitive edge.

Harman et al. (1991) show the feasibility of a distributed IR system by developing a prototype architecture and performing user testing to demonstrate usefulness. Unlike our research, which emphasizes performance, Harman et al.'s does not study efficiency issues, and they use a small text collection (i.e., less than 1GB).

Burkowski (1990) **reports** on a simulation study which measures the retrieval performance of a **distributed** IR system. The experiments explore two strategies for **distributing** a fixed workload across a small number of **servers**. The first **equally distributes** the text collection among all the servers. The second splits servers into two groups, one group for query evaluation and one group for document retrieval. This work is most closely

related to our work, but differs in several ways. He assumes a worst-case workload where each user broadcasts queries to all servers without any think time. We experiment with larger distributed configurations, vary the rate in which clients send commands, and use a more realistic user workload which uses both query and document retrieval operations.

Lin and Zhou (1993) implement a distributed IR system on a network of DEC5000 workstations using PVM (Parallel Virtual Machine) to coordinate work and network communication. They develop a new retrieval model that uses a variation of the signature file encoding scheme to map document collections over the network. Their results show large speedup improvements due to parallelization. Our research extends this work by analyzing a distributed system to increase efficiency. Since we use a simulator, we are also able to run more experiments under diverse conditions. Another advantage of our work is that we base our distributed system on a proven and effective retrieval model rather than using a new model developed to expose parallelism.

Couvreur et al. (1994) analyze the performance and cost factors of searching large text collections on parallel systems. They use simulation models to investigate three different hardware architectures and search algorithms including a mainframe system using an inverted-list IR system, a collection of RISC processors using a superimposed IR system, and a special-purpose machine architecture that uses a direct search. The focus of the work is on analyzing the trade-off between performance and cost. Their results show that the mainframe configuration is the most cost effective. They also suggest that using an inverted list algorithm on a network of workstations would be beneficial, but they are concerned about the complexity. Our research shows the effectiveness of using a network of workstations. We evaluate the performance of different configurations and explore improvements. Under realistic workloads, we present a scalable architecture for distributed information retrieval.

Hawking (1997) designs and implements a parallel information retrieval system, called PADRE97, on a collection of workstations. The basic architecture of PADRE97, which is similar to ours, contains a central process that checks for user commands and broadcasts them to the IR engines on each of the workstations. The central process also merges results before sending a final result back to the user. Hawking presents results for a single workstation and a cluster of workstations using a single 51-term query. A large-scale experiment evaluates query processing on a system with up to 64 workstations each containing a 10.2GB collection. The experiment uses four shorter queries of four to 16 terms. Our work varies several system parameters that Hawking does not, such as the query size and the rate at which clients issues queries, and we generate document commands and query commands.

Several studies have examined the use of caching in distributed IR systems (Martin and Russell 1991; Martin et al. 1990; Schatz 1990; Simpson and Alonso 1987; Tomasic and Garcia-Molina 1992). The client caches data so that operations are not repeatedly sent to the remote server. Instead, the client locally performs frequent operations. The use of caching is most beneficial for systems that are distributed over slow networks or that evaluate queries slowly. We do not study caching effects, since we implement our system on a fast local network. Furthermore, implementing a system that performs caching at the client requires a different retrieval engine.

Other researchers have investigated various data-partitioning schemes for distributed IR systems (Jeong and Omiecinski 1995; Macleod et al. 1987; Tomasic 1994; Tomasic and Garcia-Molina 1993). We model a very simple partitioning scheme in which each Inquiry server maintains an independent set of documents. The documents may be part of the same larger text collection, but we evaluate the complete query on each Inquiry server. Although we only consider one partitioning scheme, we extend previous results in several ways. Our experiments include results for both small and

large configurations. Previous research has investigated only small configurations. We investigate changes to the architecture that do not involve changes to the underlying retrieval model. Several of the partitioning schemes mentioned in the previous work require changes to the retrieval model, which possibly alters retrieval effectiveness.

The performance of our distributed system relies on the amount of parallelism we are able to exploit. Beginning in the late 1980's several researchers implemented IR systems on distributed-memory multiprocessor machines (Bailey and Hawking 1996; Cringeon et al. 1990; Frieder and Siegelmann 1991; Pogue and Willett 1987; Stanfill and Kahle 1986; Stanfill et al. 1989). We do not investigate multiprocessor systems in our experiments. Instead, we concentrate on a system in which each component of the architecture is independent. Our architecture is able to use a network of workstations instead of relying on special hardware. In a different work, we have also investigated the performance of an IR server on a shared-memory multiprocessor (Lu et al. 1998).

The TREC conference recently added the Very Large Collection track for evaluating the performance of IR systems on large text collections (Hawking and Thistlewaite 1997; Hawking et al. 1998). To handle large collections, participants use shared-memory multiprocessors or distributed architectures. The experiments in TREC-7 use 49 long queries on a 100GB collection of Web documents. The Very Large Collection track summary presents precision and query-processing time results but does not provide significant details about each system. Two of the participants present details of their distributed systems elsewhere, but neither provides significant performance evaluations (Brown and Chong 1998; Burkowski et al. 1995).

There is also a large volume of work on architectures for distributed and parallel database systems including research on performance, e.g., see Bell and Grimson (1992), DeWitt and Gray (1992), DeWitt et al. (1986), Hagmann and Ferrari (1986), Mackert and Lohman (1986), and Stonebraker et al. (1983). Although the fields of information retrieval and databases are similar, there are several distinctions which make studying the performance of IR systems unique. A major difference between database systems and information retrieval systems is structured versus unstructured data. The unstructured nature of the data in IR raises questions about how to create large, efficient architectures. Our work attempts to discover some of the factors that affect performance when searching and retrieving unstructured data. Another obvious contrast between the two systems is that relational database operators are highly parallelizable. It is unclear how to effectively parallelize natural-language queries. Furthermore, the types of common operations that are typical to database and IR systems are slightly different. For example, the basic commands in an IR system -- query evaluation and document retrieval -- differ from those in a database system. In our IR system, we are not concerned with updates (commit protocols) and concurrency control, which are important issues in distributed database systems. We assume our IR system performs updates offline.

### 3. A DISTRIBUTED INFORMATION RETRIEVAL SYSTEM

This section describes the implementation of our distributed IR system. As shown in Figure 1, the distributed system consists of a set of clients, a connection server, and a set of Inquiry servers. The distributed system enables multiple, simultaneous connections between clients and Inquiry servers. The different components of the architecture communicate using a local-area network. Each component resides on a different host and operates independently of the others. In this section, we describe the functionality and interaction between the clients, the connection server, and the Inquiry servers.

#### 3.1 Clients

The clients are lightweight processes that provide a user interface to the retrieval system. Clients interact with the distributed IR system by

connecting to the connection server, as illustrated in Figure 1. The clients initiate all work in the system, but perform very little computation. They can issue the entire range of IR commands but, in this article, we focus on query, summary information, and document retrieval commands.

A client sends query commands to the connection server. A query command consists of a set of words or phrases (terms). Either of the following occurs: the command specifies the list of Inquiry servers to search, or the client allows the connection server to determine the appropriate collections to search. In our experiments, we assume the clients choose the Inquiry servers. Query responses consist of a list of  $n$  document identifiers ranked by belief values which estimate the probability that the document satisfies the information need. The user specifies an appropriate value for  $n$ , or the system returns a default number of answers. We experiment using different values for  $n$ .

Summary information consists of the title and the first few sentences of a document. The summary information command consists of a set of document identifiers and their collection identifiers. The user may specify the number of summary entries to obtain, but, in our experiments, each client obtains summaries for 15 documents at a time. Our results show that the number of summaries has a significant effect on performance.

Clients may also retrieve complete documents by sending a document retrieval command to the connection server. The command consists of a document identifier and collection identifier. In response, the connection server returns the complete text of the document from the appropriate Inquiry server.

In our prototype distributed IR system, each client issues a command and waits for the connection server to return the results before it issues another command. Users issue query and document commands. A client automatically issues the first summary information command when it receives a query response. A client issues additional summary information and document retrieval commands at the user's request. In Section 4.3, we describe a variation of the client processing which we use to support our simulation model.

### 3.2 Connection Server

The clients and Inquiry servers communicate via a central broker called the connection server. The connection server is a lightweight process that keeps track of all the Inquiry servers, outstanding client requests, and organizes responses from Inquiry servers. The connection server continuously polls for incoming messages from clients and Inquiry servers. The connection server is able to handle outstanding requests from multiple clients.

A client sends a command to the connection server, which forwards it to the appropriate Inquiry servers. In our experiments, the command indicates the Inquiry servers. Since, each Inquiry server only processes a limited number of requests at a time, the connection server maintains a queue of outstanding requests for each Inquiry server, as illustrated in Figure 1. The connection server inserts commands from clients onto a queue if an Inquiry server is currently processing the maximum number of commands. When the connection server receives an answer from an Inquiry server, it forwards the next command on the corresponding queue to the Inquiry server.

The connection server maintains intermediate results for commands specifying multiple Inquiry servers. When Inquiry servers return results, the connection server merges them with other results. After all the Inquiry servers involved in a command return results, the connection server sends a final result to the client. Only query and summary commands may specify multiple Inquiry servers. For a query command, each Inquiry server sends its top  $n$  responses back to the connection server. The connection server maintains a sorted list of the overall top  $n$  entries until all the Inquiry servers respond. The connection server merges new results with the existing

sorted list. We assume the relative rankings between documents from independent collections are comparable, but this assumption is clearly tenuous. For example, one collection may be irrelevant to a particular query, but if the user includes it, the overall response may still include its top ranked responses. Other research is investigating techniques to automatically select appropriate collections with respect to specific queries and merge results (e.g., see Callan et al. (1995b), Voorhees et al. (1995), Viles and French (1995), Moffat and Zobel (1995), and Crowder and Nicholas (1995)). For a summary command, the connection server must read the message to determine which Inquery servers contain the appropriate documents and then send a message to each Inquery server to acquire the summary information. The connection server merges the summary information responses and sends a single message back to the client with the summary information for all the documents. The connection server does not maintain intermediate results for document retrieval commands; it simply forwards documents immediately after receiving them.

There are several benefits of using a broker to manage messages between the clients and the Inquery servers. Although the broker in our distributed system is very simple, we can enhance it to provide metadata facilities, routing, security, and more precise merging. In the context of our IR system, metadata is a concise index of the text that each Inquery server maintains. For example, the system may use the metadata to perform automatic collection selection. The connection server may also use the metadata to merge the results from each Inquery server more effectively, based on the relative importance of each collection. Routing enables a broker to manage multiple versions of the same text collection, based on availability and usage. Security becomes an issue for systems that want to protect sensitive data or allow certain clients access to certain text collections. The broker is also able to effectively manage the addition or deletion of Inquery servers. We can easily add any of these enhancements to our distributed system because a central broker manages actions between the clients and Inquery servers.

### 3.3 Inquery Servers

The Inquery server uses the Inquery retrieval engine to provide IR services such as query evaluation and document retrieval. Inquery is a probabilistic retrieval model that is based on a Bayesian inference network (Callan et al. 1992). Inquery accepts natural-language queries or structured queries. Internally, the system stores text collections using an inverted file. Previous work demonstrates that Inquery is an effective retrieval system for large, full-text databases (Callan et al. 1995a).

For this work, we use a multithreaded Inquery server that runs on a host with one CPU and one disk. We are able to achieve high resource utilization by using a multithreaded server. We use the POSIX thread library package to implement the multithreaded server. With this configuration, we have shown in related work that the best performance occurs with four threads (Lu et al. 1998). More than four threads does not yield any performance improvement, since the disk becomes overutilized. In Section 5.3.2, we briefly discuss improving the performance of our distributed system by adding more disks and threads.

Each Inquery server thread accepts a command from the connection server, processes the request, and returns the result back to the connection server. Each thread may only process one command at a time. Since we allow four threads, each Inquery server processes up to four commands at a time.

## 4. SIMULATION MODEL

In this section, we present a simulation model for exploring distributed IR system architectures. Simulation techniques provide an effective and flexible platform for analyzing large and complex distributed systems. We can quickly change the system configuration, run experiments, and analyze results without making numerous changes to large amounts of

code. Furthermore, simulation models allow us to easily create very large systems and examine their performance in a controlled environment.

To implement the simulator, we use YACSIM, a process-oriented discrete event simulation language (Jump 1993). YACSIM contains a set of data structures and library routines to manage user created processes and resources. Its process-oriented nature enables the structure of the simulator to closely reflect the actual system.

Our simulation model is simple, yet contains enough details to accurately represent the important features of the system. Section 3 describes our model's basic architecture and functionality. The hardware resources we model include CPUs, disks, and the network. We model the clients, Inquiry servers, and connection servers as different processes. Processes simulate the activities of the real system by requesting services from resources. The simulator is driven by empirical timing measurements obtained from our prototype. Section 4.2 shows that the simulator closely models the prototype implementation. We configure a simulation by defining the architecture of the distributed IR system using a simple command language. A configuration file contains the commands which the simulator reads at startup time.

Our technique for designing an environment for studying distributed information retrieval architectures is similar to the model that Brumfield et al. (1988) present. However, they simulate a distributed object-oriented database system, while our work focuses on IR systems. The main differences between database and IR systems are the type of commands and the structure of the data. IR commands include both query and document operations, and the data in IR systems are unstructured.

#### 4.1 System Measurements

To accurately model an information retrieval system, we analyze the prototype distributed Inquiry system and measure the resources used for each operation. We focus on CPU, disk, and network resources, and we do not measure memory and cache effects. Empirical measurements rather than an analytical model drive the activities performed in the simulator. The simulator uses the following parameters: query evaluation time, document/summary retrieval time, connection server time, and network latency. We obtain measurements from the prototype system using Inquiry version 3.1 running on a DEC AlphaServer 2100 5/250 clocked at 250MHz with 1024MB of memory and 2007MB of swap space. We instrument the Inquiry system to report time measurements during run time using the `ftime()` and `getrusage()` system calls. To obtain TCP network performance, we use TTCP from the US Army Research Laboratory.

We examine several different text collections and query sets to obtain system measurements. We examine TIPSTER 1, a large heterogeneous collection of full-text articles and abstracts (Harman 1992), a database containing the Congressional Record for the 103rd Congress (Croft et al. 1995), and a small collection of abstracts from Communications of the ACM (Fox 1983). The number of documents in each collection is 510,887 (1GB of text), 43,378 (500MB of text), and 3204 (2MB of text), respectively.

4.1.1 Query Evaluation Measurements. The simulator uses a simple yet accurate model to represent query evaluation time. Based on our measurements of Inquiry using our query sets, evaluation time is very strongly related to the number of terms per query and the frequency of each of the terms in the collection. Analysis of the TIPSTER 1 query set shows that there is a strong correlation of terms per query and query term frequency to evaluation time. Specifically, Pearson's correlation coefficient for both sets of data is 0.96 which indicates a strong positive relationship.(2)

Our query evaluation model is a function of the number of terms per query and the frequency of the individual query terms plus an additional overhead for the CPU and disk. Table I lists the query evaluation time values we use in our simulation. We model `eval_term_time()` as an increasing linear function of the term frequency. We also include a small CPU and disk

overhead to account for the time Inquiry spends combining the results of each of the terms. Without the overhead amount, we underestimate the evaluation time of long queries. We obtain the eval\_term\_time() function and the overhead value by using the TIPSTER I query set and text collection to measure the evaluation time for terms of different frequencies. Recall that the evaluation time has a strong positive correlation with the term frequency. We create eval\_term\_time() by fitting the term evaluation time measurements with three straight lines using a least-squares method. The first line covers term frequencies up to 10,000; the second lines covers term frequencies 10,000-100,000; and the third lines covers term frequencies above 100,000. The model is more accurate using three lines rather than a single straight line for all the data. We divide eval\_term\_time() into CPU and disk access time.

Table I. Query Evaluation Timing Values (seconds)

(MATHEMATICAL EXPRESSION NOT REPRODUCIBLE IN ASCII)

```

eval_term_time(term) = cpu_time(term) + disk_time(term)

cpu_time(term) = 0.009 + (3.2e - 6 x term) : term (is less than
or equal to) 10000

0.025 + (1.08e - 6 x term) : 10000 (is less than
or equal to) term (is less than or equal to)
100000

0.159 + (1.7e - 7 x term) : term (is greater than
or equal to) 100000

disk_time(term) = 0.061 + (4.7e - 6 x term) : term (is less than
or equal to) 10000

0.095 + (7.2e - 7 x term) : 10000 (is less than
or equal to) term (is less than or equal to)
100000

0.011 + (8.2e - 7 x term) : term (is less than
or equal to) 100000

```

Using our model, the time to evaluate a single term ranges from 0.07 seconds for a term appearing once to 0.72 seconds for a term appearing 554,568 times (the maximum term frequency in TIPSTER I after removing stopwords). Disk access time is typically larger than CPU processing time, due to the speed of the Alpha processor. Our measurements show that disk access time accounts for 32% to 90% of the total evaluation time with an average of 73%. For slower processors, CPU processing time may be larger than disk time. We notice the disk time is especially large for very small term frequencies or very large term frequencies (i.e., (is less than or equal to) 10,000 or (is greater than or equal to) 450,000).

The query evaluation model is slightly different than the model we used in previous work (Cahoon and McKinley 1996). The new eval\_term\_time() function is more accurate, and we did not include any overhead in the old model. In practice, our simulator overestimates query evaluation times, since our model assumes a query command returns the entire list of relevant answers. In our experiments, the Inquiry servers only return the top n answers. For small values of n, the actual system performs slightly less processing. We do not include this effect in our model, since it is small and difficult to characterize.

4.1.2 Document Retrieval Measurements. We measure Inquiry to determine the amount of time it takes to retrieve a document. For our text collections, the retrieval time is variable, and there is not a strong correlation between document size and retrieval time. The low correlation

is due to very quick retrieval times. In general, the sizes of the documents in our text collection are not very large, which results in fast retrieval times. In our collections, the average size of a document in the TIPSTER 1, Congressional Record, and the Fox CACM collection is 2.3KB, 11.7KB, and 0.5KB, respectively.

The simulator represents the document retrieval time for an Inquery server as a constant value, 0.027 seconds, which is the average document retrieval time for 2000 randomly selected documents from the TIPSTER 1 collection. The simulator represents retrieval time as CPU processing time plus disk access time. Our measurements show that disk access time accounts for 87% of the total retrieval time on average. To ensure our simulator is accurate for our experiments, the average document size returned by an Inquery server is the same as the average size in the TIPSTER I collection, i.e., 2.3KB. The simulator also uses the document retrieval time to compute the summary information retrieval time. We implement the summary information operation as a series of document retrieval operations. For each summary entry, an Inquery server reads a complete document. However, the Inquery servers only return the summary portion of the document. In our experiments, the average size of a document summary is 120 bytes.

4.1.3 Connection Server Time. We divide the time in the connection server into three categories; the time to encode and decode messages, the processing time for handling messages, and the time to merge results. Table II lists the formulas we use to compute the different values. The functions return seconds, except for the merge time function, which returns microseconds.

Table II. Connection Server Time Values

Message Encoding and Decoding Time (seconds)

```
query() = 0.075
summary(nsums) = 0.017 + (0.025 x nsums)
document_list(ndocs) = 0.85 + (0.0048 x ndocs)
```

Processing Time (seconds)

```
command(nservers,cmd) = 0.6 : cmd = document
                           1.3 + (0.17 x nservers) : otherwise
servers_result(cmd) = 0.7 : cmd = document
                           0.13 : otherwise
```

Merging Time (microseconds)

```
query(nanswers) = 26 + (2 x nanswers)
summary(nsums) = -75 + (24 x nsums)
```

Each time a message arrives, the connection server must decode the message to determine an appropriate action. The connection server also creates new messages and must encode the data into the message format. The time to decode a query command, query(), is a constant value. The summary(nsums) function is the time to decode and encode a summary information command; the time depends on the number of summaries, which is 15 in our experiments. The document\_list(ndocs) function is the time to decode query results (from the Inquery servers) and the time to encode the final results that the connection server sends to the client; the value depends on the number of results, or documents.

The processing time for handling messages depends on who sends the message and the message type. A command from a client is either a query, summary, or document command, which the connection server places on the queue for the appropriate destination Inquery server(s). The connection server processing depends on the type of the command. A document retrieval command is only sent to one Inquery server and takes a constant amount of

processing. A query or summary command is sent to multiple Inquery servers, so the processing time is a function of the number of servers.

Similarly, messages from the Inquery servers contain different types of results. We represent the processing time using constant values. The processing time for a document response is slightly higher than a query or summary response because the document response is larger. The connection server also spends time merging query and summary results. The time to merge query and summary results depends on the number of answers or summaries an Inquery server returns. We model the merge times as linear functions of the number of results.

4.1.4 Network Time. We represent network time as sender overhead, receiver overhead, and network latency. The sender and receiver overhead is the CPU processing time for adding and removing a message from the network. The network latency is the amount of time the message spends on the network itself. These times depend on the size of the message and the bandwidth of the network. We obtain the sender and receiver overhead times using TTCP by measuring the time to send messages between two AlphaServer 2100 5/250 workstations connected by a lightly loaded 10Mbps Ethernet. We list the equations for the sender overhead, the receiver overhead, and the network latency (each function returns its result in seconds):

```
sender_time(size) := 0.02 + (0.00003 x size)
receiver_time(size) = 0.02 + (0.00005 x size)
network_time(size) := size x 8/speed
```

In each function, size represents the number of bytes in the message, and speed is the speed of the network in bits per second.

#### 4.2 Validation

We validate the simulator's query evaluation times against the actual implementation, using a configuration consisting of a single client, a multithreaded Inquery server, and a connection server. Each of the components runs on a separate host. We do not validate the document retrieval times, since our measurements show the actual document retrieval time is variable due to the quick access times. As we mention in Section 4.1, the time to retrieve documents in the simulator is a constant value.

We validate our simulator by creating artificial queries and comparing the results from running the queries on the actual system to the results from our simulator. We randomly generate queries which correspond to the types of queries our simulator generates during our experiments (see Section 4.3). The two parameters we use to generate queries are the number of terms per query and the query term frequencies. We generate 100 short queries, 100 medium-length queries, and 100 long queries. We do not generate queries with multiple occurrences of the same term, since our model does not account for these types of queries.

Table III shows our validation results using the short, medium, and long generated queries as we vary the number of threads. Positive numbers indicate the simulator overestimates the actual system. On average, the simulator is 5.1% slower than the actual system, and the difference ranges from 3.9% faster to 27.2% slower. The difference between the actual system and the simulator tends to decrease as the number of threads increases and as the query arrival rate decreases, because we slightly overestimate the query evaluation time. In general, the simulator matches the actual system closely, but we are not able to accurately model every query. Due to our simple model it is difficult to reduce the amount of variation between our simulator and the actual system times. Given that our simulator evaluates most queries accurately, especially for four threads (average 1.2% error), and it usually conservatively overestimates, we believe it is not worth the added complexity to change the model.

Table III. Percentage Difference of Average Response Times between the Implementation and Simulator

Query Type	Arrival Rate (Lambda) per Second	Number of Threads		
		1	2	3

Short	0.5	1.5%	0.5%	-2.5%
	1	13.7%	3.0%	1.9%
	5	27.2% (a)	17.6%	3.8%
Medium	0.5	19.8%	9.0%	4.4%
	1	26.0%	12.0%	2.5%
	5	21.5%	15.7%	7.2%
Long	0.5	15.8%	6.6%	-3.8%
	1	10.1%	1.8%	-2.1%
	5	12.7%	10.3%	1.7%
Average		16.5%	8.5%	1.5%
Query Type	Arrival Rate (Lambda) per Second		Number of Threads	
		4	8	16
Short	0.5	2.5%	-0.8%	1.5%
	1	-1.5%	1.0%	-0.3%
	5	-0.4%	2.1%	1.0%
Medium	0.5	1.9%	1.2%	0.5%
	1	-3.9% (a)	1.9%	-0.4%
	5	8.8%	5.0%	4.1%
Long	0.5	-0.4%	0.4%	-0.5%
	1	1.8%	1.0%	1.7%
	5	2.4%	3.9%	3.1%
Average		1.2%	1.7%	1.2%

(a) bold

#### 4.3 Experimental Parameters

In this section, we describe the parameters we use in our simulation experiments. Table IV presents the parameters, their values, and abbreviations. We describe each parameter in more detail below.

Table IV. Experimental Parameters

Parameters	Abbrev.	Values	
Command Arrival Rate Poisson Dist. (avg. commands/sec.)	(Lambda)	0.1	0.5
Inquiry Servers	IS	1	4
Terms per Query (avg.) Shifted Neg. Binomial Dist.	TPQ	2	
Query Term Frequency Dist. from Queries	QTF	Obs. Dist.	
Answers Returned Constant Values	AR	15	
Command Mixture query:summary:document	CM	1:1:1	
Parameters		Values	
Command Arrival Rate	1	2	4
	6	8	10

Poisson Dist. (avg. commands/sec.)

Inquiry Servers	8	32	64	128
Terms per Query (avg.)	12		27	
Shifted Neg. Binomial Dist.				
Query Term Frequency		Low	High	
Dist. from Queries		Skew	Skew	
Answers Returned	100		1000	
Constant Values				
Command Mixture	1:2:2		1:3:3	1:4:4
query:summary:document				

4.3.1 Command Arrival Rate ((Lambda)). In our experiments, we model different numbers of clients by varying the rate at which we send commands to the connection server. We simulate a small or large number of clients by slowing down or speeding up the command arrival rate. The client command rate also determines the amount of "think time," i.e., the amount of time a client spends reading a document. A slow command rate simulates a large think time. We model the command rate using a Poisson distribution. Table IV lists the different distribution averages that we use in our experiments. The slowest rate averages one command every 10 seconds, and the fastest rate averages 10 commands per second.

4.3.2 Number of Inquiry Servers (IS). We experiment with both small and large system configurations. Measuring the effect of increasing the number of Inquiry servers provides insight into distributed IR architectures by identifying bottlenecks and helping to understand system utilization and scalability. We experiment with up to 128 Inquiry servers (128GB of data).

4.3.3 Terms per Query (TPQ). Table IV shows the three different average query lengths we use in our experiments. We obtain two of the values from the 103rd Congressional Record (two terms) and TIPSTER 1 query set (27 terms). We also use an intermediate value (12 terms), since the two query sets have very different characteristics. We observe that a shifted negative binomial distribution closely matches the distributions in the two query sets. The characteristics of a shifted negative binomial distribution are the number of trials,  $n$ , the probability of success,  $p$ , and the amount of shift,  $s$ . Table V shows the values we use in our experiments.

Table V. Values Used in Experiments

Description	Query Set	Length	Average		
			n	p	s
Short Queries	103rd CR	2	4	0.8	1
Medium Queries	N/A	12	2	0.77	5
Long Queries	TIPSTER 1	27	2	0.1	10

Wolfram (1992) also uses a shifted negative binomial distribution to mode) terms per query. However, we use slightly different values for  $n$ ,  $p$ , and  $s$  in order to obtain better estimates of our data sets.

4.3.4 Distribution of Terms in Queries (QTF). Zipf (1949) documented the widely accepted distribution of term frequencies in text collections. In contrast, researchers do not agree on a commonly accepted distribution for term frequencies in queries (Wolfram 1992). Figure 2 shows the distribution of our query sets. The query term frequency distributions for the query sets are similar, but the distributions do not closely match a well-known distribution. In our experiments, we use the distribution of query term frequencies from the TIPSTER 1 query set. We call this our observed query term frequency distribution. We also use a distribution that is skewed toward terms occurring less frequently and a distribution that is

skewed toward terms occurring more frequently. To obtain the low and high skewed distribution we multiply the values in the observed distribution by 0.15 and 2, respectively. Skewing the observed distribution has the effect of shifting the term frequency values in Figure 2 to either the left or the right.

(Figure 2 ILLUSTRATION OMITTED)

4.3.5 Number of Answers Returned (AR). For each query, the IR system returns a sorted list of matching documents to the clients. The list contains document identifier numbers instead of the complete text. The number of answers returned affects network traffic and processing by the connection servers. The three different constant values we use are 15, 100, and 1000 answers.

4.3.6 Command Mixture (CM). We simulate user activity by issuing query, summary information, and document retrieval commands. We vary the mixture of the commands to model different user patterns. Since we do not have statistics on user behavior, we chose four sets of values that we feel are representative. Table IV shows the different ratios we use in our experiments. For example, 1:1:1 means that the ratios of commands are the same. The 1:4:4 notation means that for each query command the client issues four summary and four document commands.

4.3.7 Simulation Output. During each simulation execution, we measure different performance statistics such as the query, summary, and document responses times. We also measure various utilization values including CPU, disk, network, connection server, and Inquiry server utilization. For each series of experiments, we display the corresponding values of the parameters and the abbreviations listed in Table IV.

## 5. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of our distributed IR architecture and present the results. Running experiments and measuring performance as a function of the command arrival rate, number of text collections, terms per query, query term frequency, answers returned, and the types of commands allows us to gain a better understanding of system performance under different workloads. To our knowledge, previous researchers have not examined the performance of a distributed IR system by varying these values. We show that each feature has a significant impact on performance, and we discuss the effect of each feature. Designers of distributed architectures for IR must account for each parameter in order to make appropriate trade-offs that maximize the performance of a particular architecture.

In Section 5.1, we show the performance when a single client searches a single text collection. We present results using the multithreaded Inquiry server (with four threads) and the single-threaded Inquiry server. As we expect, using a multithreaded server achieves better performance than the single-threaded server, due to a higher utilization of resources. The multithreaded server is able to provide reasonable response times until the command rate exceeds four or six commands/second.

In Section 5.2, we present results on an architecture that manages multiple, distinct text collections--up to 128 collections. In this case, each multithreaded Inquiry server maintains a different 1GB database, and the clients evaluate queries on a subset of the available collections. Thus, the clients potentially search up to 128GB of data. We vary the command arrival rate, the number of text collections, terms per query, query term frequency, answers returned, and the ratios of commands sent by the clients. We show that performance improves as we increase the number of Inquiry servers, up to eight or 32 servers, even though clients search more information. The improvement is due to parallelism in the summary information commands.(3)

Our results show that for short, realistic queries, several architecture configurations scale with the number of processors and degrade gracefully as the number of clients (work) increases. Our results illustrate that the system achieves good performance under varying

conditions if we can maintain a balance between connection server and Inquery server utilization. Some conditions do cause the system performance to rapidly deteriorate. For example, the architecture does not perform well for very large configurations or when either the utilization of the connection server or Inquery servers is high. We often see poor utilization for long user queries.

Varying query term frequency changes the Inquery server utilization when the Inquery servers are the bottleneck. Unfortunately, the overall system performance is still poor for large configurations. We demonstrate that varying the number of answers returned affects network utilization. Our results illustrate that increasing the number of summary and document commands relative to the number of query commands improves response time, since summary and document operations are less computationally intensive than query operations.

In Section 5.3, we change the architecture in order to improve the performance of our IR system. We show how to improve the performance by reducing the bottlenecks in the connection server and Inquery server. In Section 5.3.1, we experiment with adding connection servers to introduce more parallelism. We eliminate the connection server bottleneck for all configurations with only four connection servers. We also experiment with moving the response merging from the connection server to the clients. Unfortunately, this change does not improve performance because the amount of time spent merging is small, while the cost to send more messages decreases the benefit of moving the merging functionality. In Section 5.3.2, we describe how to improve Inquery server response times by adding hardware; we are able to increase the number of threads by adding disks. Our results show that simply adding another disk significantly improves performance when the Inquery server is a bottleneck.

Finally, in Section 5.4, we simulate a large-scale architecture that achieves high performance given our experimental parameters. Based on our results in Sections 5.2 and 5.3, we choose the parameters that provide the best results. We show that this architecture is able to provide response times under 10 seconds for short, medium, and, in some cases, long queries.

Unless otherwise stated, in each of our experiments, the clients, connection server, and Inquery servers operate as we describe in Section 3. We allocate each of the basic components in the distributed system to its own host. Each host contains its own processor, memory, and secondary storage. A local-area network with a bandwidth of 10Mbps connects the machines. Each of the Inquery servers maintains a 1GB database.

### 5.1 Base Configuration

In this section, we show the performance of a system containing one client connected to one Inquery server. The purpose of this experiment is to provide a baseline for comparison. We compare results of an Inquery server with one and four threads to illustrate the benefit of using a multithreaded server to improve response time and utilization. Previous work demonstrates that, for one disk, more threads do not improve performance (Lu et al. 1998). Thus, our base configuration uses four threads.

Table VI compares the command response times for an Inquery server using 1 and 4 threads. As the command rate increases, the response times for the multithreaded server become much better than the single-threaded server. We first see a large difference between 1 and 4 threads when the command rate increases from 2 to 4 commands/second. For example, Table VI shows that the query response time increases by a factor of 31.2 when the command rate increases from 2 to 4 commands/second. When using 4 threads, we see an increase of only 2.35. The results show that the multithreaded server always achieves response times under 1 minute. However, there are large increases in response times once the command rate exceeds 4 commands/second. We see a factor of 16.7 increase in query response time when the command rate increases from 4 to 6. The table also shows that the document response times become very similar to the query and summary

command times as the command rate increases. We explain each of these effects by showing the utilization values of the Inquery servers as the command rate increases.

Table VI. Command Response Times in Base Configuration (seconds)

Command	Number of Threads	Command Rate (Average per Second)				
		0.1	0.5	1	2	4
Query	1	0.49	0.52	0.58	0.87	27.15
	4	0.49	0.50	0.53	0.65	1.53
Summary	1	0.44	0.48	0.54	0.81	28.02
	4	0.44	0.47	0.50	0.61	1.57
Document	1	0.05	0.08	0.16	0.43	27.17
	4	0.04	0.08	0.13	0.27	1.24

  

Command	Number of Threads	Command Rate (Average per Second)		
		6	8	10
Query	1	68.24	89.01	104.53
	4	25.60	44.20	57.20
Summary	1	68.70	89.32	105.02
	4	26.02	44.78	57.66
Document	1	68.08	88.97	104.66
	4	25.51	44.19	57.33

Figures 3 and 4 show utilization values using the single-threaded and multithreaded Inquery server, respectively. The figures show the utilization of the server process, the disk, and the CPU.(4) For a single thread, the utilization of the server reaches the maximum value after four commands/ second. Disk usage accounts for most of the time in these configurations. Figure 4 illustrates that disk utilization approaches 100%, which limits efficiency. For example, we see that CPU utilization reaches a maximum of 36% because the Inquery server waits for the disk and is unable to fully utilize the CPU. Results from other experiments show that the maximum CPU utilization is approximately 36% regardless of query length. When users issue longer queries, CPU and disk utilization reach their maximums more quickly.

(Figures 3-4 ILLUSTRATION OMITTED)

## 5.2 Searching Multiple Text Collections

In this section, we measure the performance of a distributed IR system that maintains multiple 1GB text collections. In this configuration, each client selects a random subset of the available collections to search and searches half of the available collections on average. Each collection has an equal chance of being chosen. The workload increases both as a function of the number of Inquery servers and the command arrival rate (i.e., the total amount of information to search increases with the number of Inquery servers). This workload mimics the scenario when the connection server automatically selects an appropriate subset of the available collections to search. It also models the situation when the user manually chooses a subset of the collections to search.

In the following sections, we present results showing the effect of our experimental parameters on the performance of the architecture. These parameters include the command arrival rate, number of text collections, terms per query, query term frequency, answers returned, and number of query, summary, and document commands. We list the values for each

parameter in a table accompanying each of the figures which appear in this section.

5.2.1 Effect of Terms per Query. In this section, we discuss the performance effect of using short, medium, and long queries. We see that query length has a dramatic impact on system performance. The results in this section show that bottlenecks sometimes occur in the connection server and the Inquiry server. In general, system performance is poor when users issue medium or long queries. However, user queries tend to be very short, and our architecture is able to efficiently handle many configurations.

5.2.1.1 Short Queries (TPQ = 2). Figure 5 shows the query response time when clients issue short queries. We vary the number of Inquiry servers, and we vary the command rate. Figure 5 illustrates an improvement in query response time as we add Inquiry servers and exploit parallelism, up to eight or 32 Inquiry servers.(5) Going from one to eight Inquiry servers improves performance by a factor of 42, when clients issue 10 commands/second. However, when the system contains more than eight or 32 Inquiry servers, performance begins to degrade. This result is interesting because the amount of text to search increases as the number of Inquiry servers increases. For example, when the number of Inquiry servers doubles, a client potentially searches twice as much information.

(Figure 5 ILLUSTRATION OMITTED)

More detailed measurements reveal the improvement stems from increased parallelism during summary information retrieval. Recall that a single summary information operation retrieves 15 summaries. A system with one Inquiry server contains all the documents on the same machine. However, a system with multiple Inquiry servers distributes the documents among the available Inquiry servers. The 15 summary entries may reside on different Inquiry servers resulting in parallel access of the summary information. In the best case, each of the 15 entries is located on different Inquiry servers.

As Figure 5 shows, the response time degrades very slowly as we increase the command rate for eight or 32 Inquiry servers. For example, as we increase the command rate by a factor of 100 (0.1 to 10 command/second), query response time degrades by a factor of 2.6 for eight or 32 Inquiry servers. For the other combinations of Inquiry servers, the response time rapidly increases as the command rate increases. Bottlenecks in either the connection server or the Inquiry server cause this decrease in performance.

Figures 6 and 7 show the utilization of the connection server and Inquiry server when clients issue short queries. The graphs show opposite effects; as the command rate increases, connection server utilization increases, and Inquiry server utilization decreases. The system achieves the best performance when neither utilization values are high. It is apparent that as the system size increases the connection server becomes a bottleneck, causing performance to degrade. The problem is due to the large number of messages sent to the connection server. Although the connection server processes messages quickly, it is unable to keep up with the quantity of messages. The maximum number of results sent to the connection server from the Inquiry servers at any single point in time is Inquiry servers

x  
threads

(Figures 6-7 ILLUSTRATION OMITTED)

The first two rows of Table VII show the response times for query, summary, and document commands for short queries when the system contains eight and 128 Inquiry servers. Using short queries, the architecture achieves the best response times using eight Inquiry servers. For example, the largest query response time is 1.4 seconds when clients average 10 commands/second, which is quite reasonable. We also show the response times when the system contains 128 Inquiry servers to compare the results for medium and long queries. Using 128 Inquiry servers, performance begins to

degrade after four commands/seconds. Table VII also illustrates that the summary and document response times grow at a similar rate as the query commands. This trend occurs in each of the experiments we perform. Thus, we will only report query response times throughout the rest of the article.

Table VII. Response Times for Short, Medium, and Long Queries (seconds)

Query Length	Command	Command Rate			
		0.1	0.5	1	2
Short (8 IS)	Query	0.5	0.5	0.6	0.6
	Summary	0.2	0.2	0.3	0.3
	Document	0.04	0.05	0.07	0.1
Short (128 IS)	Query	1.1	1.2	1.4	1.9
	Summary	0.1	0.2	0.4	0.9
	Document	0.0	0.1	0.3	0.7
Medium (128 IS)	Query	3.8	4.6	5.8	11.1(a)
	Summary	0.3	1.4	3.0	8.1(a)
	Document	0.2	0.8	1.7	4.3(a)
Long (128 IS)	Query	9.3(a)	15.6	41.9	192.1
	Summary	1.2	7.7(a)	30.8	170.6
	Document	0.7	3.7(a)	15.3	128.1
Query Length	Command	Command Rate			
		4	6	8	10
Short (8 IS)	Query	0.7	0.8	1.0	1.4(a)
	Summary	0.4	0.6	0.8	1.1(a)
	Document	0.2	0.3	0.47	0.67(a)
Short (128 IS)	Query	7.0(a)	39.2	62.1	76.3
	Summary	6.1(a)	37.6	59.4	72.5
	Document	5.3(a)	34.4	54.6	66.2
Medium (128 IS)	Query	54.0	94.1	110.8	119.9
	Summary	46.7	86.4	103.7	112.1
	Document	31.5	68.9	85.7	94.4
Long (128 IS)	Query	304.7	342.6	387.9	399.9
	Summary	280.8	320.5	360.6	373.7
	Document	236.4	274.4	311.0	322.5

5.2.1.2 Medium Queries (TPQ = 12). System performance degrades when clients issue medium-length queries. Unlike the situation when clients issue short queries, the response times do not initially improve and then degrade after eight Inquiry servers. Instead, response times improve as we add Inquiry servers, and the system achieves the best performance with 128 Inquiry servers. The best response time when clients issue 10 commands/second is 119.9 seconds, which is larger than the best time for short queries by a factor of 86. However, the third row of Table VII shows that the system achieves a response time of 11 seconds or less with a command rate less than two commands/second. The reason for the poor performance when clients issue commands quickly is that the Inquiry servers are unable to process commands fast enough. The main bottleneck in the Inquiry server is the disk, which reaches 100% utilization after the command rate exceeds two commands/second.

5.2.1.3 Long Queries (TPQ = 27). When clients issue long queries, the response times become poor once clients average greater than one command

every two seconds. As we increase the command rate from 0.1 to 10 on a system with 128 Inquiry servers, the performance degrades by a factor of 43 (compare this with factors of 2.8 and 31.5 for short and medium queries, respectively). The poor performance is due to the high utilization of the Inquiry server's disk. The bottom row of Table VII shows that performance is noticeably worse when clients issue medium-length queries. For example, when clients issue two commands/second, the response time for long queries increases by a factor of 17.3 and 320 over medium and short queries, respectively. Similar to the situation with medium queries, the system performance improves as the number of Inquiry servers increases. Unfortunately, the response times are poor, except when clients issue commands at an extremely slow rate.

5.2.2 Effect of Query Term Frequency. Varying the query term frequency only has a noticeable effect on performance when the Inquiry servers are the bottleneck. The reason is that queries containing terms which occur more frequently in the text collection evaluate more slowly. The opposite is true of queries containing terms which occur less frequently. Our results reflect this difference in evaluation time. We see little, if any, changes in performance due to the query term frequency when the connection server is the bottleneck.

In Figures 8 and 9, we present the query response times for low and high skewed query term frequencies when clients issue short queries. As we expect, the figures show that we obtain better performance for low skewed queries and worse performance for high skewed queries when we compare the results to the observed distribution (see Figure 5). Compared to the results in Figure 5, using low skewed terms decreases the query response time by 36% to 58% in a configuration with eight Inquiry servers. In the same configuration, using high skewed terms increases query response time by 24% to 121%. The performance differences are larger as the command rate increases. Although we do not show graphs for medium and long queries, our experiments show similar performance trends as the results in Figures 8 and 9. Low skewed queries improve performance, and high skewed queries degrade performance. Unfortunately, the performance is still quite poor.

(Figures 8-9 ILLUSTRATION OMITTED)

We do not see any differences in performance once the connection server becomes the bottleneck. In configurations involving 64 or 128 Inquiry servers, the connection server is the bottleneck when clients issue commands at a relatively fast rate (see Figure 6). In this situation, the response times when using the low, observed, or high distribution are similar. For example, when clients issue 10 commands/second, the query response times in a configuration with 64 Inquiry servers for the low, observed, and high distributions are 14, 16, and 16.3 seconds, respectively. In a configuration with 128 Inquiry servers, the query response times are 74.5 (low), 76.3 (observed), and 78 (high).

5.2.3 Effect of Answers Returned. Changing the number of answers returned for each query has several direct effects on system performance. As we decrease the number of answers, we see an improvement in the command response times, especially when there are a large number of Inquiry servers. Another effect is that the amount of processing the connection server performs is related to the number of answers returned. Finally, we observe that the number of answers affects network usage. Our results show that reducing the number of answers returned is an effective technique for improving performance, especially when the connection server is the bottleneck. Figures 10 and 11 show the query response times when the Inquiry servers compute the top 15 and 100 answers for each query, respectively. When we compare these figures to the results in Figure 5 for 1000 answers, we see the number of answers an Inquiry server computes has an impact on the performance of several configurations. Notice the change in the x-axis for Figures 10 and 11 (i.e., 0 to 10 versus 0 to 80 in Figure 5). Although all cases show poor performance when there is only one Inquiry server, the graphs show that the query response time improves as we

decrease the number of answers returned. There is a large improvement for 64 and 128 Inquiry servers when we decrease the number of answers from 1000 to 100. For example, when clients issue 10 commands/second the response time improves by 92% (64 IS) and 96% (128 IS). Reducing the number of answers to 15 also improves performance although the improvement between 100 to 15 is small for most configurations.

(Figures 10-11 ILLUSTRATION OMITTED)

Our results suggest that the response time improvements are due to lower utilization of the connection server and network. Figure 6 shows the connection server utilization when the Inquiry servers return 1000 answers. The utilization values for 64 and 128 Inquiry servers becomes extremely high when the command rate increases (e.g., the utilization reaches 100%). The maximum utilization values when Inquiry servers return 100 and 15 answers is 82% and 70%, respectively.

Figure 12 shows the relative network utilization when clients request 15, 100, and 1000 answers. The relative values are the averages of all the clients. The error bars show the lowest and highest values. Figure 12 shows that as the number of clients increases, the number of answers returned has a relatively large impact on network utilization. As we reduce the number of answers, the network utilization decreases. For example, the network utilization for 15 and 100 answers is 31% and 36% of the network utilization for 1000 answers, respectively, when there is one Inquiry server. As the number of Inquiry servers increases, the gap between 1000 answers and 15 or 100 answers becomes larger.

(Figure 12 ILLUSTRATION OMITTED)

In terms of absolute values, network utilization becomes quite large as the number of Inquiry servers and the command rate increases. The maximum network utilization when 1000 answers are returned is 93%, which occurs for 128 Inquiry servers when the command rate exceeds six commands/second. However, for a small number of Inquiry servers or a relatively slow command rate, the network utilization is very low and is usually less than 20%. As a comparison, the highest network utilization when Inquiry servers return 15 answers is 7%.

5.2.4 Effect of Changing the Command Mixture. In this section, we vary the command mixture to include more summary and document commands. We run tests using ratios of 1:1:1, 1:2:2, 1:3:3, and 1:4:4 (the number of query:summary:document commands). The performance of the system improves as the number of summary and document commands increases. We expect an improvement, since it takes less resources to service a summary or document command compared to a query.

Figures 13 and 14 illustrate the effect of increasing the number of summary and document operations on response time for an architecture with eight and 128 Inquiry servers, respectively. Our results show significant improvements in performance as the mixture of summary and document commands increases. The most impressive improvements occur when there are a large number of Inquiry servers. When clients issue 10 commands/second and the command ratios change from 1:1:1 to 1:4:4, the query response time decreases by 46% for eight Inquiry servers and 95% for 128 Inquiry servers. For 128 Inquiry servers, we also see large improvements when the command mixture changes from 1:1:1 to 1:2:2.

(Figures 13-14 ILLUSTRATION OMITTED)

Our system also achieves reasonable response times when clients issue medium-length queries and the command mixture is 1:4:4. We observe response times less than 20 seconds for 64 and 128 Inquiry servers (compare this with 120 seconds in Table VII). Unfortunately, increasing the number of summary and document commands does not produce reasonable response times when clients issue long queries.

### 5.3 Improving Performance

In this section, we discuss changes to the basic architecture to improve system performance. Our results show that the system scales for short queries up to a certain point; if we add too many Inquiry servers,

then performance degrades, since the connection server becomes a bottleneck. In this section, we show that adding a small number of connection servers eliminates the bottleneck in the connection server. We have also shown that, in some cases, the Inquery server is the bottleneck due to high disk utilization. In this section, we improve the performance of the Inquery server by adding more disks and threads to alleviate the bottleneck.

We have also run experiments trying to improve the performance of the connection server by moving the merging functionality to the clients. Recall that the connection server is responsible for collecting and merging intermediate results before sending the final answer to the client. We do not present the results because moving merging functionality has very little effect on performance. We do not see significant changes in performance because the merge time is fast, and the time to send more messages from the connection server to the client offsets the time the connection server previously spent merging results.

5.3.1 Improving Connection Server Performance. The results in Section 5.2 show that the connection server is a bottleneck when clients issue short queries and the system contains a large number of Inquery servers. The cause of the bottleneck is that the connection server is unable to handle requests quickly enough. Previous results show the connection server is the bottleneck more frequently when the architecture uses a slower processor (Cahoon and McKinley 1996). We can eliminate this bottleneck using two or four connection servers to handle up to 128 Inquery servers. Although using a faster processor for the connection server is one solution, we show that performance significantly improves using a small number of connection servers. Simply increasing the speed of the CPU does not completely solve the problem because we anticipate that the functionality of the connection servers will increase in the future which may offset gains in processor speeds.

When the system has multiple connection servers, the clients evenly divide among the connection servers, and each connection server maintains a link to all the Inquery servers. In the basic architecture, the connection server maintains a queue of outstanding requests for each Inquery server. If the Inquery server is busy, the connection server adds the request to the queue. In the multiple-connection-server system, the connection server immediately forwards requests to the Inquery servers. Each of the Inquery servers maintains its own queue of outstanding requests. We have run experiments which indicate that simply moving the queues to the Inquery servers does not significantly change performance.

In these experiments, we also increase the speed of the network from 10Mbps to 100Mbps. Initial results show that the network becomes a bottleneck when we increase the number of connection servers. The network utilization approaches 100% in configurations with 64 or 128 Inquery servers. Due to the high network utilization, we see very little benefit when we add connection servers. In order to show the benefit of additional connection servers, we increase the network speed. If it is not possible to increase the network speed, then we can decrease network utilization by decreasing the number of answers returned (see Section 5.2.3). In the single-connection-server architecture, increasing the speed of the network has very little benefit. Figures 15 and 16 compare the query response times when the architecture has one, two, or four connection servers for 64 and 128 Inquery servers. We only show results when the connection server is a bottleneck. There is little, if any, benefit of increasing the number of connection servers when there is no bottleneck.

(Figures 15-16 ILLUSTRATION OMITTED)

The figures show relative performance values compared to the absolute times from Figure 5 when the architecture has only one connection server. The figures show that increasing the number of connection servers significantly improves the response time. When the architecture has 64 Inquery servers, the difference in the normalized response time increases

as the command rate increases. For example, Figure 15 shows that the normalized response time at 10 commands/second is 10% and 8% for two and four connection servers. These relative values translate into absolute times that are 1.67 and 1.32 seconds (compared with 16 seconds for the one-connection-server case).

When the architecture has 128 Inquiry servers (Figure 16), the performance using two connection servers improves until the command rate reaches six commands/second. At that point, the connection servers start to become a bottleneck and the response time increases. However, when the architecture has four connection servers, the relative response time continues to improve as the command rate increases. When clients issue 10 commands/second, the relative response times for two and four connection servers are 29.5% and 2.9%, respectively. The absolute response times are 22.5 and 2.2 seconds (compare with 76 seconds for the one-connectionserver configuration).

**5.3.2 Improving Inquiry Server Performance.** The Inquiry server is a bottleneck when the command rate is high and when there are a small number of Inquiry servers. Recall that each Inquiry server allows up to four threads and runs on a host with a single CPU and disk. Our results show the disk is the primary problem. We see disk utilization values as high as 98.5%. In previous work, we have investigated the performance of an information retrieval engine on symmetric multiprocessors that contain multiple CPUs and disks (Lu et al. 1998). The results suggest that we can improve performance by adding resources. In this section, we illustrate how adding only one more disk significantly improves performance when the Inquiry servers are a bottleneck. We also increase the number of threads from four to eight for this experiment. It is possible to improve performance even further by adding CPUs, more disks, and increasing the number of threads, but the focus of this article is the distributed architecture and not the peak performance of a multiprocessor system.

Figure 17 shows the results of adding another disk to each host on query response time. We see significant performance improvements when the Inquiry servers are the bottleneck. In a configuration with 1 Inquiry server, we see improvements by a factor of 15.84, 3.68, and 2.66 when clients issue 6, 8, and 10 commands/second, respectively. When the issue rate is less than 6 commands/second, the response times are very similar. We also see improvements in a configuration with 4 and 8 Inquiry servers when clients issue 10 commands/second. When the number of Inquiry servers is greater than 8, including more disks has no benefit because the Inquiry servers are not the bottleneck in those configurations.

(Figure 17 ILLUSTRATION OMITTED)

Figure 18 suggests that one reason for the performance improvement is due to better utilization of the Inquiry server. Compare the utilization values in Figure 18 to Figure 7. The largest utilization when the Inquiry servers contain two disks is 260% out of 300% (compare to 140% out of 200% when the Inquiry servers contain a single disk).(6) The high utilization comes from an increase in the CPU utilization, which approaches 70% coupled with just a 4% decrease in disk utilization (per disk).

(Figure 18 ILLUSTRATION OMITTED)

#### 5.4 Achieving High Performance in Large Systems

Ultimately, users of IR systems want to enter queries and obtain fast and accurate results. They are not concerned with the different factors that affect performance. In contrast, system designers must account for many parameters in order to maximize performance, especially when creating large architectures. In this section, we simulate a high-performance architecture in order to determine the best possible performance in a highly utilized, large-scale system given our experimental parameters.

We simulate an architecture containing four connection servers and 128 Inquiry servers (each with 1GB of data). Each Inquiry server contains one CPU and two disks and allows up to eight threads. We vary the query length (TPQ) and the command arrival rate ((Lambda)). We set the other

parameters to the values that produce the best results. These parameters include the query term frequency (QTF = Low Skew), number of answers returned (AR = 15), and command mixture (CM = 1:4:4).

Table VIII shows the query response times for our high-performance architecture (we italicize values greater than 10 seconds). For short queries, the query response time is very fast, less than 0.46 seconds, and does not significantly increase as the command rate increases. The architecture also achieves response times under seven seconds when users issue medium-length queries. Using long queries, we see response times under 10 seconds until the command rate exceeds two commands/second. Even when the command rate is 10 commands/second, the response time is under one minute.

Table VIII. Best Query Response Times for Large Systems (seconds)

IS	QTF	AR	CM	Command Rate (Average per Second)					
128	Low	15	1:4:4						
Query Length									
Short	0.41	0.42	0.42	0.41	0.42	0.43	0.46	0.46	
Medium	1.88	2.06	2.22	2.23	2.76	3.73	4.15	6.57	
Long	4.74	5.38	6.31	7.51	13.71	25.98	41.98	57.77	

The response times in Table VIII occur only in best-case scenarios. Unfortunately, system designers do not have complete control over each of the experimental parameters. The system designer controls the number of Inquiry servers and connection servers, the amount of hardware resources, and the number of answers returned. User behavior defines the other parameters, which include the number terms per query, the query term frequency, and the command mixture. However, these results suggest it is possible to provide high performance under many different realistic conditions. The effectiveness of the system depends on architectural decisions as well as typical user behavior.

#### 6. SUMMARY

To keep pace with the increasing amounts of online information, the performance of information retrieval systems must improve. In this article, we present an implementation of a distributed IR system to achieve coordinated, concurrent, and scalable access. We develop a flexible simulation model to examine the performance of the prototype using a wide variety of parameters, workloads, and configurations. We present results that measure system response time, utilization, and identify bottlenecks as a function of several key IR characteristics: client command rate, number of text collections, number of terms per query, query term frequency, number of answers returned, and the mixture of query, summary, and document commands.

We present results for a system that distributes multiple text collections over a set of independent hosts. The best performance occurs for architectures with either eight or 32 Inquiry servers when the system is able to exploit parallelism in the summary information commands. Performance degrades for more than 32 Inquiry servers when the connection server becomes a bottleneck, especially for short queries. Using these results, we change the architecture to improve performance and scalability by eliminating bottlenecks that occur in either the connection server or the Inquiry servers. By adding a small number of connection servers to coordinate a large number of clients and Inquiry servers, the system maintains scalable performance at higher workloads. We also illustrate that including another disk significantly improves Inquiry server performance in cases when the disk is a bottleneck.

Our work presents an efficient and scalable architecture for large-scale distributed IR. We model and measure a large architecture, up to 128GB of data on 128 machines, and vary key IR features in order to provide insight into designing distributed IR systems. We also illustrate that it is important to model both query and document commands because the heterogeneity of the commands has a significant impact on performance.

#### ACKNOWLEDGMENTS

We would like to thank Bob Cook and Kathleen Dibella for help with the development of the prototype system. We also thank Jamie Callan and Bruce Croft for their contributions to this work.

(1) We have examined architectures that distribute a single text collection among the Inquiry servers in previous work (Cahoon and McKinley 1996).

(2) Pearson's correlation coefficient ranges between -1 and +1. Values -1 and +1 indicate a perfect linear relationship and occur when all points lie on a downward or upward sloping line, respectively. A value of 0 indicates there is no linear relationship.

(3) In previous work we have also run experiments when partitioning a single 1GB text collection among all the Inquiry servers (Cahoon and McKinley 1996). The results are very similar to those in Section 5.2.

(4) The Inquiry server utilization is equal to the disk plus the CPU utilization values. In our system with one disk and one CPU, the maximum utilization of the Inquiry server process is 200%.

(5) The response times for 32 Inquiry servers are only slightly higher than eight Inquiry servers.

(6) On a system with two disks and one CPU, the maximum potential utilization is 300%, which means that the two disks and the CPU are always working concurrently.

#### REFERENCES

BAILEY, P. AND HAWKING, D. 1996. A parallel architecture for query processing over a terabyte of text. Tech. Rep. TR-CS-96-04. Department of Computer Science, Australian National Univ., Canberra, Australia.

BELL, D. AND GRIMSON, J. 1992. Distributed Database Systems. Addison-Wesley, Reading, MA.

BROWN, E. W. AND CHONG, H.A. 1998. The GURU system in TREC-7. In Proceedings of the 7th Text Retrieval Conference (TREC-7),

BRUMFIELD, J. A., MILLER, J. L., AND CHOU, H. T. 1988. Performance modeling of distributed object-oriented database systems. In Proceedings of the International Symposium on Databases in Parallel and Distributed Systems (Austin, TX, Dec. 5-7), J. E. Urban, Ed. IEEE Computer Society Press, Los Alamitos, CA, 22-32.

BURKOWSKI, F. J. 1990. Retrieval performance of a distributed text database utilizing a parallel process document server. In Proceedings of the International Symposium on Databases in Parallel and Distributed Systems (Dublin, Ireland, July), 71-79.

BURKOWSKI, F., CORMACK, G., CLARKE, C., AND GOOD, R. 1995. A global search architecture. Tech. Rep. CS-95-12. Computer Science Dept., University of Waterloo, Waterloo, Canada.

CAHOON, B. AND MCKINLEY, K. S. 1996. Performance evaluation of a distributed architecture for information retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96, Zurich, Switzerland, Aug. 18-22), H.-P. Frei, D. Harman, P. Schable, and R. Wilkinson, Eds. ACM Press, New York, NY, 110-118.

CALLAN, J. P., CROFT, W. B., AND BROGLIO, J. 1995a. TREC and TIPSTER experiments with INQUERY. Inf. Process. Manage. 31, 3 (May-June), 327-343.

CALLAN, J. P., CROFT, W. B., AND HARDING, S. M. 1992. The INQUERY retrieval system. In Proceedings of the 3rd International Conference on Database and Expert System Applications (Valencia, Spain, Sept.),

CALLAN, J. P., LU, Z., AND CROFT, W. B. 1995b. Searching distributed collections with inference networks. In Proceedings of the 18th Annual

International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95, Seattle, WA, July 9-13), E. A. Fox, P. Ingwersen, and R. Fidel, Eds. ACM Press, New York, NY, 21-28.

COUVREUR, T. R., BENZEL, R. N., MILLER, S. F., ZEITLER, D. N., LEE, D. L., SINGHAL, M., SHIVARATRI, N., AND WONG, W. Y. P. 1994. An analysis of performance and cost factors in searching large text databases using parallel search systems. *J. Am. Soc. Inf. Sci.* 45, 7 (Aug.), 443-464.

CRINGEAN, J. K., ENGLAND, R., MANSON, G. A., AND WILLETT, P. 1990. Parallel text searching in serial files using a processor farm. In *.Proceedings of the 13th International Conference on Research and Development in Information Retrieval (SIGIR '90, Brussels, Belgium, Sept. 5-7)*, J.-L. Vidick, Ed. ACM Press, New York, NY, 429-453.

CROFT, W. B., COOK, R., AND WILDER, D. 1995. Providing government information on the Internet: Experiences with THOMAS. In *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries (DL '95, Austin, TX, June)*, 19-24.

CROWDER, G. AND NICHOLAS, C. 1995. An approach to large scale distributed information systems using statistical properties of text to guide agent search. In *Proceedings of the CIKM Workshop on Intelligent Information Agents (Baltimore, MD, Dec.)*,

DEWITT, D. AND GRAY, J. 1992. Parallel database systems: the future of high performance database systems. *Commun. ACM* 35, 6 (June), 85-98.

DEWITT, D., GRAEFE, G., KUMAR, K. B., GERBER, R. H., HEYTENS, M. L., AND MURALIKRISHNA, M. 1986. GAMMA--A high performance dataflow database machine. In *Proceedings of the 12th International Conference on Very Large Data Bases (Kyoto, Japan, Aug.)*, VLDB Endowment, Berkeley, CA.

Fox, E. A. 1983. Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. *Tech. Rep. 83-561*. Cornell University, Ithaca, NY.

FRIEDER, O. AND SIEGELMANN, H. T. 1991. On the allocation of documents in multiprocessor information retrieval systems. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '91, Chicago, IL, Oct. 13-16)*, E. Fox, Ed. ACM Press, New York, NY, 230-239.

HAGMANN, R. B. AND FERRARI, D. 1986. Performance analysis of several back-end database architectures. *ACM Trans. Database Syst.* 11, 1 (Mar. 1986), 1-26.

HARMAN, D. K., Ed. 1992. *Proceedings of the 1st Text Retrieval Conference*. (TREC-1, Gaithersburg, MD, Nov.). National Institute of Standards and Technology, Gaithersburg, MD. NIST Special Publication 500-217.

HARMAN, D., MCCOY, W., TOENSE, R., AND CANDELA, G. 1991. Prototyping a distributed information retrieval system that uses statistical ranking. *Inf. Process. Manage.* 27, 5 (1991), 449-459.

HAWKING, D. 1997. Scalable text retrieval for large digital libraries. In *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries (Pisa, Italy, Sept.)*, Springer Lecture Notes in Computer Science, vol. 1324. Springer-Verlag, Vienna, Austria.

HAWKING, D. AND THISTLEWAITE, P. 1997. Overview of the TREC-6 very large collection track. In *Proceedings of the 6th Text Retrieval Conference (TREC-6, Nov.)*, E. Voorhees and D. Harman, Eds.

HAWKING, D., CRASWELL, N., AND THISTLEWAITE, P. 1998. Overview of TREC-7 very large collection track. In *Proceedings of the 7th Text Retrieval Conference (TREC-7)*,

JEONG, B.-S. AND OMIECINSKI, E. 1995. Inverted file partitioning schemes in multiple disk systems. *IEEE Trans. Parallel Distrib. Syst.* 6, 2 (Feb.), 142-153.

JUMP, J. R. 1993. YACSIM reference manual. Version 2.1.1. Rice University, Houston, TX.

LIN, Z. AND ZHOU, S. 1993. Parallelizing I/O intensive applications

for a workstation cluster: a case study. *SIGARCH Comput. Arch. News* 21, 5 (Dec. 1993), 15-22.

LU, Z., MCKINLEY, K. S., AND CAHOON, B. 1998. The hardware/software balancing act for information retrieval on symmetric multiprocessors. In *Proceedings of Euro-Par'98* (Sept., Southampton, UK),

MACKERT, L. F. AND LOHMAN, G. M. 1986. R\* optimizer validation and performance evaluation for distributed queries. In *Proceedings of the 12th International Conference on Very Large Data Bases* (Kyoto, Japan, Aug.), VLDB Endowment, Berkeley, CA, 149-159.

MACLEOD, I. A., MARTIN, P., NORDIN, B., AND PHILLIPS, J. R. 1987. Strategies for building distributed information retrieval systems. *Inf. Process. Manage.* 23, 6 (July 1, 1987), 511-528.

MARTIN, T. P. AND RUSSELL, J. I. 1991. Data caching strategies for distributed full text retrieval systems. *Inf. Syst.* 16, 1 (1991), 1-11.

MARTIN, T. P., MACLEOD, I. A., RUSSELL, J. I., LEESE, K., AND FOSTER, B. 1990. A case study of caching strategies for a distributed full text retrieval system. *Inf. Process. Manage.* 26, 2 (1990), 227-247.

MOFFAT, A. AND ZOBEL, J. 1995. Information retrieval systems for large document collections. In *Proceedings of the 3rd Text Retrieval Conference (TREC-3)*, D. Harman, Ed. National Institute of Standards and Technology, Gaithersburg, MD, 500-525.

POGUE, C. A. AND WILLETT P. 1987. Use of text signatures for document retrieval in a highly parallel environment. *Parallel Comput.* 4, 3 (June), 259-268.

SCHATZ, B. R. 1990. Interactive retrieval in information spaces distributed across a wide-area network. TR 90-35. Department of Computer Science, University of Arizona, Tucson, AZ.

SIMPSON, P. AND ALONSO, R. 1987. Data caching in information retrieval systems. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '87, New Orleans, LA, June 3-5)*, C. T. Yu and C. J. Van Rijsbergen, Eds. ACM Press, New York, NY, 296-305.

STANFILL, C. AND KAHLE, B. 1986. Parallel free-text search on the connection machine system. *Commun. ACM* 29, 12 (Dec. 1986), 1229-1239.

STANFILL, C., THAU, R., AND WALTZ, D. 1989. A parallel indexed algorithm for information retrieval. In *Proceedings of the 12th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '89, Cambridge, MA, June 25-28)*, N. J. Belkin and C. J. van Rijsbergen, Eds. ACM Press, New York, NY, 88-97.

STONEBRAKER, M., WOODFILL, J., RANSTROM, J., KALASH, J., ARNOLD, K., AND ANDERSON, E. 1983. Performance analysis of distributed data base systems. In *Proceedings of the 3rd Symposium on Reliability in Distributed Software and Database Systems* (Clearwater Beach, FL, Oct.),

TOMASIC, A. S. 1994. Distributed queries and incremental updates in information retrieval systems. Ph.D. Dissertation. Department of Computer Science, Princeton Univ., Princeton, NJ.

TOMASIC, A. AND GARCIA-MOLINA, H. 1992. Caching and database scaling in distributed shared-nothing information retrieval systems. Tech. Rep. STAN-CS-92-1456. Stanford University, Stanford, CA.

TOMASIC, A. AND GARCIA-MOLINA, H. 1993. Performance of inverted indices in distributed text document retrieval systems. In *Proceedings of the 2nd International Conference on Parallel and Distributed Systems* (Dec.), 8-17.

VILES, C. L. AND FRENCH, J. C. 1995. Dissemination of collection wide information in a distributed information retrieval system. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95, Seattle, WA, July 9-13)*, E. A. Fox, P. Ingwersen, and R. Fidel, Eds. ACM Press, New York, NY, 12-20.

VOORHEES, E. M., GUPTA, N. K., AND JOHNSON-LAIRD, B. 1995. Learning collection fusion strategies. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval (SIGIR '95, Seattle, WA, July 9-13), E. A. Fox, P. Ingwersen, and R. Fidel, Eds. ACM Press, New York, NY, 172-179.

WOLFRAM, D. 1992. Applying informetric characteristics of databases to IR system file design, Part I: informetric models. *Inf. Process. Manage.* 28, 1 (Jan.), 121-133.

ZIPF, G. K. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Reading, MA.

Received: April 1997; revised: February 1998, February 1999, and June 1999; accepted: September 1999

This material is based on work supported by the National Science Foundation, Library of Congress, and Department of Commerce under cooperative agreement number EEC-9209623. This work is supported in part by NSF Multimedia award CDA-9502639. Kathryn S. McKinley is supported by an NSF CAREER Award CCR-9624209. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsors.

Authors' addresses: B. Cahoon and K. S. McKinley, Department of Computer Science, University of Massachusetts, Amherst, MA 01003; email: cahoon@cs.umass.edu; mckinley@cs.umass.edu; Z. Lu, Village Networks, 100 Village Court, Hazlet, NJ 07730; email: zlu@vill.com.

Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

(C) 2000 ACM 1046-8188/00/0100-0001 \$5.00

Categories and Subject Descriptors: C.2.4 (Computer-Communication Networks): Distributed Systems; C.4 (Computer Systems Organization): Performance of Systems -- Performance attributes; H.3.4 (Information Storage and Retrieval): Systems and Software

General Terms: Experimentation, Performance

Additional Key Words and Phrases: Distributed information retrieval architectures

BRENDON CAHOON and KATHRYN S. MCKINLEY University of Massachusetts and ZHIHONG LU Village Networks

COPYRIGHT 2000 Association for Computing Machinery, Inc.

DESCRIPTORS: Information storage and retrieval systems--Design and construction; Network architecture--Design and construction; Client/server architecture--Design and construction

GEOGRAPHIC CODES/NAMES: 1USA United States

FILE SEGMENT: AI File 88

?